THE DEVELOPMENT AND TESTING OF A STRUCTURED TRAINER'S REPORT FOR SUMMATIVE ASSESSMENT IN GENERAL PRACTICE

Peter Neil JOHNSON

THESIS FOR DOCTOR OF MEDICINE DEGREE

Submitted to The Department of Postgraduate Medical Education, University of Warwick.

Based on work undertaken at the Department of Public Health and Primary Care, University of Oxford, 1994-7.

March, 1999

CONTENTS

TABLES, FIGURES AND APPENDICES

TABLES

FIGURES

APPENDICES

ACKNOWLEDGEMENTS

I am most grateful to all those who have had some involvement in this project. I recognise that without all the support I have received it would not have been possible to design the project, to undertake the studies involved, and to submit this thesis. However, there are a number of people to whom I should like to pay particular tribute.

I am deeply indebted to the Department of Health for their very substantial financial support for this project. In particular I should like to highlight the complete absence of any attempts to influence this project in any way.

I should like to record particular thanks to Dr John Hasler who was of enormous help, both when the original idea was being formulated and refined and throughout the project; his ability to facilitate motivation, clear thinking and scholarly reporting provided inspirational support throughout this project.

Dr John Toby and Professor Janet Grant also provided enormous support as part of the steering group for this project; their ability to criticise ideas and reports in a constructive way was of particular value.

I should like to record my thanks to Dr John Wilmot and Dr Bob Strachan for acting as my supervisors for this doctorate; in particular, their willingness to provide time for meetings and to read draft reports (and, in particular, draft versions of this thesis) was very welcome. I am also most grateful to Professor Jeremy Dale for his support in the latter stages of developing this thesis. Additional support was also provided by Dr Roger Gadsby, Dr Ala Szcepura and Dr Mike Graveney

I am very grateful to Professors Robert Burgess and John Bligh for examining and providing detailed advice on the thesis .

I would like to record my thanks to my partners at the Medical Centre, Shipston-on-Stour (Graham Williams, Theo Schofield, Chris Thorogood, Caroline Nixon, Jane Gilder and David Williams) for allowing me to have protected time for this project.

I am grateful to all those who work in the Postgraduate Office at the Medical School in Oxford (in particular Barbara Vaughan, Joan Watts and Gill Pattinson) for helping to coordinate the financial arrangements and steering group meetings.

I should like to record my considerable feelings of gratitude to all those who were willing to be involved in the various studies; in particular I should like to record thanks to the following people: all the regional advisers who were so supportive to the project and were willing to supply information about the trainers in their regions; all the trainers in the thirteen groups who allowed me to join their groups and interview them; all the trainers who responded to the questionnaire regarding the contents of the report form; all those who gave up their time to attend the consensus conference to set standards (Dr P Lane, Dr M McGhee, Dr A Barker, Dr A Robinson, Dr C Halliday, Dr C Matheson, Dr G MacKinnon, Dr J Schofield, Dr R King, Wing Cdr S Law, Lt Col G Lawrenson, Dr A Rogers, Dr B Ormston, Dr J den Bak, Dr R Bethel, Dr G Boulos, Dr A Membrey, Dr I McLean, Dr S Lazar, Dr J Oubridge, Dr R Milne, Dr W Patterson, Dr M Rhodes, Dr A Dunn, Dr A Brzezicki, Dr T Vell, Dr K Emerson, Dr J Allen, Dr A Watson, Dr A Wilkinson) and to those who took part in the subsequent consultation exercise; all those doctors who had recently completed their vocational training who completed

questionnaires to look at the content validity of the report; to all those trainers (Drs Priestman and Dean, Hassey and Wilkinson, Hartley-Brewer and Cottee, Hughes and Vernon, Maxmin and Masters, D Mercer and J Mercer, Little and Miller, Parrish and Norris, Paul and Walters, Cox and McDermott, Purce and Rea, Rowlands and Butcher, Peppiatt and Corbett, Poll and Law, M Rogers and A Rogers, Moy and Angus, Benett and Greenaway, McGill and Hodgson, Tew and Deakin, Buchan and Skinner, Denny and Allan, Gallagher and Ball, Blick and Levitt, Lees and Bee, Tinkler and Williams, Hinton and Jones, Stott and Cowlard, Merchant and Hall, Wilkes and Walker, Keeling and Flynn, Gilmore and Baird, House and Diack, Curson and Pietroni, Kingsland and Thomas, Mulka and Vaughan, Robinson and Merry, Harris and Hasenfuss, Browne and Brant, Martin and Ross, Middleton and Hanlon, Grace and Godby, Lloyd-Smith and Dowling, Redmond and Redmond, Shaw and Duncan, Smith and Patton, Syme and Cunningham, Dummer and Sutherland, Wookey and Shapley, McCann and Osborne, Stone and Webster, Cassels and Coughlin, Jenkins and Reid, Robertson and Hamilton, Jenns and Kent, Strachan, Leigh and Knott), and all their trainees, who undertook the field testing of the trainer's report; and finally to those doctors, whose identities are unknown to me, who provided considerable help in clarifying the issues in their role as peer reviewers for the journals to which the reports from this project were submitted.

I leave my greatest thanks until last; my wife, Helen, has been hugely supportive of me undertaking this project despite the considerable sacrifices that this has meant to our time together - without her support none of this would have been possible. Similarly, my children Bryony and Thomas have learnt to accept that much of my free time has been devoted to completing both the project and the thesis - I am grateful for their understanding!

DECLARATION

As each of the studies reported in this thesis was completed they were submitted to peer-reviewed journals for consideration for publication. This has resulted in the following publications:

1. Johnson N, Hasler J, Toby J, Grant J. The contents of a trainer's report for summative assessment in general practice: the views of trainers. Br J Gen Pract 1996; 46: 135-139.
2. Johnson N, Hasler J, Toby J, Grant J. Consensus minimum standards for use in a trainer's report for summative assessment. Br J Gen Pract 1996; 46: 140-144.
3. Johnson N, Hasler J. Content validity of a trainer's report for summative assessment in general practice. Med Educ 1997; 31: 287-292.
4. Johnson N, Hasler J, Toby J, Grant J. Pilot testing of a structured trainer's report for summative assessment in general practice. Education for General Practice 1997; 8: 308-315.

The multiple authorship of these papers acknowledges the role of the steering group in developing the project and discussing the presentation of the results. However the actual studies described in this thesis and these papers were all undertaken solely by myself.

SUMMARY

The central theme of this thesis is the place of a report provided by the trainer on the performance of the trainee as part of a process of regulating entry to independent general medical practice in the United Kingdom (summative assessment). The thesis aims both to analyse the place of a such a report within a process of summative assessment and to consider whether it is possible to develop a report form for this purpose that enables aspects of the general practitioner trainee's skills, knowledge, attitudes and practice to be assessed by the trainer in a feasible, valid and reliable way.

It is argued that the certification process for entry to independent general practice in the United Kingdom needs review and that tests of performance, such as a trainer's report, have a particular role in such a process; that such tests should be criterion-referenced; and that a number of properties are of particular importance in the development and testing of a trainer's report in the context of the assessment of doctors completing general practitioner training in the United Kingdom.

A set of research objectives are delineated for a sequential series of five research studies. Using a variety of methods (semi-structured group interviews, postal questionnaire surveys, consensus conference, and pilot testing), these studies demonstrate: that there is a specific place for a trainer's report; that valid contents can be selected and minimum standards set; that the report form that has been developed is reliable and feasible and allows discrimination; and that, should it be widely adopted, there is a strong need for further testing, a continuing quality assurance system and further developmental work.

It is concluded that summative assessment does have a role in providing an initial step in assuring the public of the quality of doctors entering independent general practice and that the report form developed here is suitable for wide application within such a process. It is also reasoned that a number of lessons about the application of such a process, and the inclusion of such a report, in other settings can be learnt. In particular it is suggested that a report provided by a trainer may have a particular role in assessment when the requirement is the assessment of performance of complex attributes within the context of training designed to enable the trainee to carry out a particular purpose but that it should rarely be used as the sole instrument.

CHAPTER ONE - INTRODUCTION

1.1 Background

In 1992 the body responsible for regulating entry into general medical practice in the United Kingdom (the Joint Committee on Postgraduate Training for General Practice (JCPTGP)) announced a plan to introduce a system of assessment designed to ensure that doctors completing training for general medical practice were fit to practise independently (Joint Committee on Postgraduate Training for General Practice Working Party on Assessment, 1992a). It was suggested that this system, known as "summative assessment", should consist of four components one of which would be a report provided by the general practitioner trainer on the performance of the doctor-in-training.

The proposal makes three assumptions: that a summative assessment process will perform a useful function, that a report provided by the trainer is a desirable method as part of such an assessment process, and that it is possible to design a report form that will enable accurate assessment. The aim of this thesis is to examine these three assumptions. Firstly the thesis aims to analyse the role of a summative assessment process and the place of a trainer's report within it. Secondly, through the research component, a consideration is made of whether it is possible to design a report form to be completed by a trainer that does enable feasible, valid and reliable assessment of the performance of a doctor training within the context of general medical practice in the United Kingdom.

The research took place between mid-1994 and the end of 1996. When the research project started the exact proposals for summative assessment were not widely known,

and trainers were largely unaware of the move to have a summative assessment process. By the time the research was concluded, the requirement of trainers to submit the trainee to an assessment based on a report form designed for the purpose was widely understood. This required a rapid shift in the way that trainers were expected to think and behave. Consequently the research was conducted on a background of continuous political negotiation within the profession, and aspects of the discussion of the results will reflect this.

Summative assessment for doctors completing training for general medical practice in the United Kingdom was introduced on a professionally-led basis on the 4th September, 1996 and became mandatory from 31st January, 1998 (Anonymous, 1997).

1.2 A guide to the thesis

Following this introductory chapter there are five main components to the thesis.

In the first component, chapter two, an analysis of the place of summative assessment within training is made with particular reference to the application of such a process to general medical practice in the United Kingdom. The chapter considers the question "why is summative assessment needed?".

The second component, chapter three, examines more specifically the place of a trainer's report within such a process. Three issues are examined: "if there is to be summative assessment what form should such a process take?"; "if there is to be a trainer's report within summative assessment, what are the technical requirements of such an assessment

instrument?"; and "if these technical requirements are to be met, what specific research questions need to, and can, be answered?".

The third component, chapters four and five, contains the research component of the thesis. Chapter four considers methodological issues for the research, and concludes with the detail of the methods selected for each study. Chapter five describes the results of the studies, in particular focusing on the degree to which desired technical requirements for a trainer's report have been met.

The fourth section, chapter six, draws conclusions from the results of the research through an analysis of the main results of the research, the methodological issues that affect the weight that can be placed on the research findings, the principal issues that arise from the results, and an examination of the place of these findings within the broader context of work previously published. A discussion of the areas in which research and continuing development are still needed is also included.

In the final component, chapter seven, the wider issues raised in the first component of this thesis are reconsidered. In particular an attempt is made to generalise the messages that arise from the analysis undertaken and research findings made in this thesis by a consideration of three questions. The first question returns to the themes discussed in chapter two, namely: "is this summative assessment process likely to address the concerns which forced its introduction?". This is followed by two more general questions: "how does this summative assessment process help us in our thinking about assessment within education?" and "how does this trainer's report help us in our thinking about such assessment instruments?".

1.3 Glossary of terms

Below is a list of the definitions of a number of terms used in this thesis.

Assessment instrument/method: this refers to an individual technique for measuring an attribute of the individual being assessed.

Assessment system/process: this refers to a combination of instruments which, together, provide a mechanism for the assessment of a broad range of attributes.

Assessor: the individual undertaking the assessment.

Assessee: the individual being assessed.

Trainee: the individual being trained.

Trainer: the individual training others.

GP Trainee (Registrar): a doctor training in a general practice for a career in general medical practice in the United Kingdom.

Professional self-regulation: the process by which a professional group determines whether or not individuals should become, or remain, members of that profession.

CHAPTER TWO - BACKGROUND: THE RATIONALE FOR A SUMMATIVE ASSESSMENT PROCESS IN GENERAL PRACTICE

2.1 Introduction

The proposal for the introduction of a summative assessment process at the point of entry to independent general medical practice in the United Kingdom (Joint Committee on Postgraduate Training for General Practice Working Party on Assessment, 1992a) rests upon the assumption that such a process will perform an important function. In this chapter this basic assumption is examined - it considers the question "why have a process of summative assessment?".

Firstly, the definition of summative assessment is clarified. Secondly, an analysis is made of the forces driving the introduction of such a process within general practice and the current position. Thirdly, an analysis of the problems inherent in such an assessment process is made. The chapter concludes by assessing whether the weight of arguments support or refute the development of summative assessment in general practice.

2.2 Definitions - what is summative assessment?

The term assessment is derived from the Latin verb *adsidere*, the translation of which is "to sit beside someone" (Anonymous, 1989). Rowntree develops this notion by describing assessment as "an attempt to know that person", going on to define assessment within an educational setting as "occurring whenever one person, in some kind of interaction, direct or indirect, with another, is conscious of obtaining and

interpreting information about the knowledge and understanding or abilities and attitudes of that other person" (Rowntree, 1977). Bloom adds a predictive component, defining assessment as "the act of gathering and processing evidence about human behaviour under given conditions for purposes of understanding, predicting, and controlling future human behaviour" (Bloom, 1968). From these definitions I would highlight two particular aspects of assessment: firstly, that the purpose of assessment is to know about particular attributes of an individual in some detail; secondly that assessment requires both the gathering of evidence and an interpretation to be made of the meaning of that evidence.

A distinction needs to be made between assessment and a term that is sometimes used synonymously - namely evaluation. Rowntree states "if assessment tries to discover what the student is becoming or has accomplished, then evaluation tries to do the same for a course or learning experience or episode of teaching" (Rowntree, 1977); that is, evaluation is concerned with knowing the strengths and weaknesses of an educational intervention (or institution), whilst assessment is concerned with the strengths and weaknesses of the individual.

Within education, numerous authorities (Bloom et al. 1971; Rowntree, 1977; Black and Devine, 1986) have divided assessment into two forms - namely, formative (or pedagogic) and summative (or classificatory) assessment. In formative assessment the assessment activity is aimed at informing the teacher and learner about what further learning is needed - the purpose is to "form or alter the course of study for each student" (Haile, 1977) through a "diagnostic" approach (Black and Devine, 1986); the emphasis is on development. In summative assessment the assessment activity is focused on the

"final overall impact of the instructional sequence" (Haile, 1977); the emphasis is on achievement. A large range of possible purposes for both of these forms of assessment have been defined (Cronbach, 1964; Klug, 1974; Rowntree, 1977; Broadfoot, 1979). The principal purpose for which summative assessment is used is the "selection" of individuals who are likely to perform adequately in the future (Cronbach 1964; Broadfoot, 1979; Desforges, 1989), thereby "maintaining the standards" of the group which the individual is to join (Rowntree, 1977). This process may include the awarding of a certificate or license, and may therefore be termed licensure or certification (Black, 1993).

In the setting of medical education and training, summative assessment has become synonymous with a process of assessment of the individual, timed to coincide with the end of training, in which the practice of an individual doctor is assessed as a means of ensuring that minimum standards of practice have been reached prior to entry to the profession (Joint Committee on Postgraduate Training for General Practice Working Party on Assessment, 1992a). In this way the process is being used as a process of certification; the process is obligatory (or "imperative" (Black and Devine, 1986)) rather than optional.

2.3 The purpose: why have summative assessment in general medical practice in the United Kingdom?

It is crucial to clarify why summative assessment might be desirable in general medical practice in the United Kingdom. In particular, what is/are the problems that a summative assessment process is designed to solve? This analysis considers two issues: what forces

are driving the introduction of more rigorous assessment, and to what extent does the current system meet these expectations?

2.3.1. Forces driving assessment

Five types of force appear to be driving the development of a summative assessment process for general practitioners in the United Kingdom. These are political forces, societal forces, educational forces, international forces and professional forces.

Political forces

The strongest drive seems to arise, perhaps unsurprisingly, from politicians - as Raven suggests "competence in modern society (is) ..... primarily dependent on political competence" (Raven, 1991); that is, definitions of professional competence will be highly dependent on the political will of the time.

There has undoubtedly been in recent years a considerable political drive to improve both quality and accountability throughout the National Health Service. In 1989 the Prime Minister (the Rt. Hon. Margaret Thatcher MP) stated clearly the desire of the U.K. Government of the time for the National Health Service to become more directly answerable to patients - "the government's desire for a health service that gives patients more choice and rights as consumers" (Thatcher, 1989). Subsequent governments have focused particularly on the issue of quality within the National Health Service, including the need for standards to be maintained in the medical profession; as the Rt. Hon. Stephen Dorrell MP (Secretary of State for Health, 1995-7) stated 'there is no higher priority in the Health Service than the maintenance and development of professional standards.....Patients and the public have a right to expect that doctors'

professional performance is of the highest quality' (Hibbs, 1995). The signals are clear - the government is demanding accountability and high standards.

At the same time the profile of primary care, traditionally performing in the shadow of the secondary sector, is being enhanced. Successive governments have shown a determination to have a National Health Service led by primary care (NHS Executive, 1996; Department of Health (England), 1996; Department of Health (England), 1997). Although one ostensible reason given for this is that "family doctors (are) regarded by the government as the jewel in the crown of the NHS" (Fletcher, 1996) it is probable that a major reason for focusing closely on primary care is a belief that, through its role in gatekeeping access to expensive secondary care services, primary care could be an important tool in the desire to control expenditure on health care.

The combination of these two drives means that there is now considerable political pressure to ensure that the clinical standards of doctors working in primary care are of a high standard (NHS Executive, 1996).

Societal forces

Does this political will reflect the concerns of the people who will use the service? Two forms of evidence are available - evidence about the views of groups of patients on their doctors, and evidence about changes occurring in the relationship between society and its general practitioners.

Whilst much of the evidence from patients has focused on practices (rather than individuals) as deliverers of a service (Baker and Streatfield, 1995; Baker, 1996), there is

a limited body of evidence concerning the views of patients about their doctors. In a study that considered the views of consumers about general practitioners in the United Kingdom Williams and Calnan demonstrated that the key dimensions in maintaining the satisfaction of the public with their general practitioners were "communication", "the relationship", and the "professional skills" of the general practitioner (Williams and Calnan, 1991). Similarly, in a review of consumer satisfaction with primary care in four European countries, Calnan et al. demonstrated that the nature of the patient-doctor relationship and the professional skills of the doctors were key to consumer satisfaction in all four countries (Calnan et al. 1994); stepwise regression analysis of their results suggested that medical skills were judged to be second in importance only to the amount of information that doctors gave (which, in itself, might be considered to be part of the communication skills of the doctor). On a global scale, Wensing et al. have recently undertaken a systematic review of the literature from around the world on patient priorities for general practice care (Wensing et al. 1998). Their findings confirm that patients are particularly concerned about "informativeness", "humaneness" and "competence/accuracy". In summary, there is strong evidence to suggest that patients are concerned about the clinical and communication skills of their general practitioners.

The evidence on the relationship between society and its general practitioners suggests that the doctor-patient relationship in primary care has moved considerably from one of paternalism to one that is more consistently based on partnership. Cartwright undertook extensive interviews with patients and their general practitioners in 1964 (Cartwright, 1967) and again in 1977 (Cartwright and Anderson, 1981). From the first study she concluded that people appeared to select their doctor is a "casual" way; she also declared that "people .... probably realise that they cannot assess (the doctor's) professional

competence" (Cartwright, 1967). In her subsequent study she found that 30% of patients over the age of 25 and 57% of doctors believed that in the previous ten years patients had become more likely to question whether their doctor was right (Cartwright and Anderson, 1981). This change embraced a number of areas of care including clinical skills and judgement, and communication skills. She found few other major changes between the findings of the two studies, thus suggesting that this change had occurred despite the fact that most change in the relationship between society and its doctors appears to occur at a much slower rate. This change in the relationship has subsequently been borne out in a number of publications (Zeitlyn, 1979; Stocking, 1991; Doyle, 1996; Kee, 1996; Quality and Consumers Branch, 1996). The consequence of such a change is that patients are more questioning about the ability of their doctors to perform and less willing to accept it on trust alone; whether this has driven, or been driven by, the politicians, the inevitable result is again one of increasing emphasis on quality and accountability.

Educational forces

The third driving force is educational. This argument stems from the view that education and assessment are closely intertwined - assessment ensures that the educational process is working. Assessments provide information both on individuals who have undertaken their education within that system and on the education system as a whole. A number of authors have highlighted the closeness of this relationship (Kandel, 1936; Rowntree, 1977) arguing that, just as there should not be an assessment process without an educational process to support it, so there should not be an educational process without an associated assessment process to provide a measure of institutional and individual success.

Currently there does seem to be political pressure to target education to deliver trainees fit for a particular purpose rather than seeing education purely as an acceptable academic exercise in its own right (Hickox, 1995); the corollary of such a drive is that the associated assessment processes must enable assessment of fitness for the particular purpose at which the educational process is targeted.

International forces

The international support for assessment has three components. Firstly there is now a vast international experience of assessment methods (e.g. for assessment in a medical setting see Fabb and Marshall, 1983 or Anonymous, 1994); it is not tenable to argue that there can be no assessment through lack of experience. Secondly, the general drive for greater accountability within education is not confined to the United Kingdom (Broadfoot, 1979). Thirdly, the increasing movement of people between countries (particularly between the United Kingdom and Europe) requires qualifications to be transferable (defined, for medical education, in a European Council Directive (Anonymous, 1993)); consequently it must be possible to scrutinise the basis on which qualifications are granted to analyse their acceptability in other countries.

Professional forces

Pressure from within the profession has arisen in two ways. Firstly, Gray has argued that education for general practice must, in common with education elsewhere, be based on a paradigm one component of which is assessment (Gray, 1977). Secondly, concern about a small number of trainees not being of the required standard was openly expressed some years ago (Royal College of General Practitioners, 1985). To address these concerns the

body controlling entry to general medical practice in the U.K. (the JCPTGP) established some years ago that the certificate issued to indicate "satisfactory completion" of training should be considered to indicate "satisfactory performance" (Irvine et al. 1990). The JCPTGP has developed this view further and has recognised the need for the introduction of a formal system of assessment if the credibility of the profession is to be maintained (Joint Committee on Postgraduate Training for General Practice Working Party on Assessment, 1992a) and has been the principal professional body driving its introduction.

2.3.2 The current system of regulation

Regulation of the medical profession

Within the medical profession in the United Kingdom the regulatory system that has developed has been based on the principle of professional self-regulation - that is, it is the professionals themselves who regulate entry to the profession.

The earliest recorded professional qualifying examinations were those for the medical profession. Established in 1815 by one of the professional groupings within medicine their function was "to determine the competence and limit access to membership of the profession" (Gipps and Stobart, 1993). A statutory framework for the regulation of entry to the medical profession was initiated by the Medical Act of 1858 (Anonymous, 1858); it was this act that established the General Medical Council as a professional self-regulatory body for the medical profession. A number of reviews (Ministry of Health and Department of Health for Scotland, 1944; Royal Commission on Medical Education, 1968; United Kingdom Parliament House of Commons Social Services, 1975) and Acts

of Parliament (Anonymous, 1886; Anonymous, 1978; Anonymous, 1983) have established a position in which the General Medical Council clearly recognises the need for training beyond that provided by undergraduate medical courses for those who seek to practise independently. Consequently, if adequate regulation is to be achieved through a guarantee of the knowledge and skills necessary for independent practice, assessment placed solely at the completion of undergraduate training is insufficient - further assessment during or at the end of such a period of extended training is required.

The principal rationale for vesting regulation within the professions themselves is that each profession has a "specialised body of knowledge...which the average citizen cannot fully comprehend" (Cruess and Cruess, 1997) and that, as a consequence, only the professionals themselves can undertake regulation. Nevertheless, self-regulation does pose considerable problems - in particular the problem of how experts are to be held accountable to non-experts. Self-regulation is predicated on an assumption that "professionals will put the welfare of both the patient and society above their own" (Cruess and Cruess, 1997). But, as Gold enquires "how (can) the principle that decisions should be made by those most affected be reconciled with the principle that decisions should be made by those with experience and training in the area" (Gold, 1981). Consequently, as the current President of the General Medical Council points out, the independence that a profession gains by the allowance of self-regulation must be considered to be a privilege and not necessarily a right - if the profession fails to regulate itself properly it would be reasonable for society to demand that the privilege of independence was removed (Irvine, 1997). It is therefore essential that public confidence in the system is preserved; the public must be able to trust the profession "to undertake

proper regulatory action when individuals do not perform competently or ethically" (Irvine, 1997).

What does this mean in the late twentieth century - should regulation of the professions remain with the professions or should it be replaced? I believe that there are two arguments which suggest that self-regulation, with appropriate safeguards, should remain. The first is a practical reason - in reality, what is the alternative to self-regulation? The alternative would be some form of external social control (Johnson, 1972), through some form of review body. Such a body would still need to command the respect both of the public (and their politicians) and the profession. Because it is quite probable that their regulatory methods would need to be very similar, the only substantial difference between such a body and a self-regulatory body would be one of public and professional perception. If this is true then the corollary is that, provided that public and professional expectations are maintained in an appropriately balanced way, a professional self-regulatory body could still undertake the function of regulation. In other words, with adequate safeguards, a system of professional regulation may provide an acceptable regulatory function. The second argument is a more profound sociological one. Johnson argues that professions use the power of "mystification" to increase their social distance from their clients - "the greater the social distance, the greater the helplessness of the client" (Johnson, 1972); he concludes by arguing that the greater the helplessness, the greater the risk is of exposure of the client to exploitation by the professional, and the greater the need for social control of the professions. If he is correct, then the corollary must surely be that the less the helplessness, the less the need for social control. If, as was argued on p.24, paternalism is being replaced by a greater

sense of doctor-patient partnership, the need for social control rather than self-regulation may be lessening.

Regulation within general medical practice

Following extensive lobbying (College of General Practitioners, 1966) the Royal Commission on Medical Education recommended that a specific period of training should precede independent general medical practice (Royal Commission on Medical Education, 1968); this is now known as vocational training. As a consequence training schemes were first developed in a number of areas in the nineteen-seventies (Gray, 1992). The regulatory framework for vocational training was finally enshrined in the NHS (Vocational Training) Regulations of 1979 (Anonymous, 1979) which resulted in vocational training becoming compulsory for those doctors entering general practice after 1981. As a consequence doctors wishing to enter general practice have to undertake a minimum of three years of post-registration experience, of which at least one year must take place in a general practice approved for the purposes of vocational training. The outcome of this regulation has been the establishment of a strong educational system, with over three thousand general practitioners being approved as trainers, supported in turn by Course Organisers (who coordinate training between the hospital posts and general practice posts); ultimate control of the vocational training lies with the Directors of Postgraduate General Practice Education (formerly known as Regional Advisers) working in conjunction with the Postgraduate Deans.

Unfortunately, this development has not been supported by the development of adequate assessment; there remains no nationally accepted compulsory end-point assessment within general practice. The regulation of entry to general practice in the United

Kingdom is controlled by the JCPTGP, a professional self-regulatory body formed of representatives of the two main bodies representing the interests of general practitioners in the United Kingdom - the Royal College of General Practitioners and the General Medical Services Committee of the British Medical Association. This body issues certificates indicating competence in the specialty of general practice; it acts as the "U.K. Competent Authority" for general practice (Anonymous, 1993). Certification is based on the collection of an acceptable combination of certificates. At the end of each component of training, the supervising doctor (hospital consultant or general practitioner) is asked to provide a "certificate of satisfactory completion" of the post (the so-called VTR1 form for general practice training posts and VTR2 form for hospital posts). Questions about the validity of this system can be raised. The rate of failure of certification is extremely low, the most recent failure rates being in the region of 0.26% (Joint Committee on Postgraduate Training for General Practice, 1992; Joint Committee on Postgraduate Training for General Practice, 1993; Joint Committee on Postgraduate Training for General Practice, 1994), a failure rate that must call into question the validity of the system - can it really be true that virtually all doctors (399 in every 400) completing vocational training are competent? The question of validity has been further reinforced by the findings of a pilot study of a summative assessment process which resulted in a failure rate closer to 5% (Campbell and Murray, 1996). Whilst the reasons for this low failure rate have not been formally investigated, I would suggest that there are at least four problems with this current system. The first two are to do with the process of assessment - it is not clear as to what areas of performance should be assessed (i.e. the "content" of the assessment), nor is it clear what standards are being used for the assessment (i.e. the "referencing" of the assessment). The second two are to do with the

role of the assessor - namely the risk of a doctor being signed up without any observation of actual performance taking place, and the risk of collusion between trainer and trainee.

Many doctors currently completing vocational training do submit themselves to a voluntary end-point assessment which exists in the form of the membership examination of the Royal College of General Practitioners. This examination could not be considered to be an adequate summative assessment process for four reasons. Firstly, the Royal College of General Practitioners alone does not have a professional mandate to act as the regulatory body. Secondly, the examination is voluntary; it is not taken by all doctors and so can not be considered to safeguard all members of the public. Thirdly, the aim of the examination is to support excellence (Haslam, 1998) rather than to ensure minimum standards; consequently, because the standard is high, there may be a risk that a substantial body of doctors whose performance is above a minimum standard but who fall short of excellence will fail the test thereby undermining its credibility. Fourthly, it is not comprehensive; the test does not include any test of the clinical performance of the doctor. Consequently the membership examination is very limited as an assessment instrument for regulatory purposes.

Summary

In the introduction to this chapter, the question was posed: "what is/are the problems that a summative assessment process is designed to solve?". In summary I would identify three particular problems that a summative assessment process could help to solve. These are: *social credibility* - doctors need to be able to demonstrate that those entering the profession are of an adequate standard to undertake the service required by society and its politicians; *professional credibility* - the assessment process needs to

command the respect of those who will be subject to it - with adequate safeguards, a summative assessment process led by the profession could ensure social credibility whilst also commanding the respect of the profession; *educational credibility* - a sophisticated educational system must be supported by an adequate assessment process.

The current regulatory system both for the medical profession as a whole and for general medical practice in particular, does not adequately address these issues. The pursuit of a new system of assessment at the completion of vocational training is therefore justified. What are the arguments against such a process?

2.4 The arguments against summative assessment

Rowntree has suggested a number of potential negative effects of assessment (Rowntree, 1977). Of his extensive list, the following components seem to be of particular importance in the context of an assessment process designed for the selection/deselection of doctors for independent practice:

- stereotyping - assessment is affected by the previous knowledge of the assessee; the assessor notices most those features that correspond to an initial diagnosis (the "selective perception" effect).
- observation effects - the certainty of the conclusions drawn from observations of performance is reduced because the act of observing alters the performance (either for the better or for the worse); this effect may be even more pronounced when the assessee knows what the trainer already thinks about the assessee's performance (the performance starts to follow as a "self-fulfilling prophecy").
- extrinsic rewards - whilst Rowntree is concerned mainly with the assessee being more interested in obtaining the certificate than the learning, for a summative assessment

process focusing on deselection it is the converse effect that is likely to pose the greater problem - namely that the assessee is more concerned with avoiding failure than with learning; this effect may vary according to the testing method used - for example a test of knowledge based on multiple-choice questions may drive the assessee to concentrate on the techniques required for such answer formats rather than on the knowledge required.

- bureaucracy effects - a system that is highly bureaucratic may produce failures purely as a result of failing to overcome the bureaucratic hurdles, rather than true failures. A system that is highly bureaucratic may also cause resistance within those who will be implementing it.

Gipps and Stobart identify another major potential side-effect - the so-called "curriculum backwash effect" (Gipps and Stobart, 1993). This is the effect of the assessment on the curriculum being taught; for a national assessment method the result would be that the curricula of training programmes throughout the country would be geared more to ensuring passes in the examination and less to the learning required to develop good doctors. It is the institutional form of the "extrinsic rewards" side-effect listed above.

Summary

It can not be assumed that the introduction of a system of summative assessment would be an unqualified success. Any new assessment process will produce some side-effects. Crucially, a judgement will have to be made about whether the benefits of such a system outweigh, or are outweighed by, these problems.

2.5 Conclusions

The arguments put forward in the first section of this chapter suggest that the drive for public accountability of the medical profession, and general practice in particular, is strong. Despite a sophisticated educational system, the profession can not currently be considered to be adequately protecting the public at the point of entry to general practice. If the profession wishes to retain self-regulation, rather than find itself subject to external social control, the profession will need to find an answer.

Summative assessment as a process for regulating entry to general practice looks attractive. Nevertheless, there are risks with such an approach. Do the potential benefits outweigh the potential risks? My view is that continuation of the current system is no longer tenable. Although certain, predictable, risks do exist, I believe that the risks to the credibility of the medical profession, and general practice in particular, through a failure to address the limitations of the current system are significantly greater. Indeed, because the predictable risks are only risks to members of the profession, the use of such risks as a rationale for resisting change may well serve only to increase a belief that the profession is failing to act to provide adequate protection for patients.

Consequently I believe that there is considerable merit, and some urgency, to a consideration of the development of an acceptable summative assessment process.

Chapter three builds on the conclusion that the introduction of a system of summative assessment for general medical practice is desirable through an analysis of the practical issues involved - in particular: what form should summative assessment for general practice take?; if there is to be a trainer's report, what are the requirements of such a

report?; and if a new trainer's report is needed, what are the research questions that need to be answered?

CHAPTER THREE - BACKGROUND: ASSESSMENT METHODS AND THE REQUIREMENTS OF A TRAINER'S REPORT

In chapter two it has been concluded that, within the setting of general medical practice in the United Kingdom, a number of forces are driving a review of the system of regulation of entry to the profession. In particular, a number of legitimate expectations are not being adequately met by the current system. It is concluded that the proposal of the JCPTGP to introduce a system of summative assessment prior to entry to independent practice is justified. It is now necessary to consider how this assessment might be undertaken.

The purpose of this chapter is to examine what methods might be used for such a summative assessment process with a particular focus on the potential role of a report provided on the performance of the trainee by the trainer.

To do this, three issues are examined. Firstly, from a range of possible assessment methodologies, which are likely to be the most suitable methods and how should they be selected? This section begins by reviewing general principles which are then applied to the particular context of doctors entering independent general medical practice. Secondly, if a trainer's report is considered to be a suitable assessment method in this particular context, what are the technical requirements of such a report? This section begins with an examination of the general principles followed by an examination of existing experience of the use of such reports. Thirdly, if such a report is to be

developed *de novo* for this particular purpose, what are the research questions that need to be answered? This section begins with an examination of how the technical requirements for assessment instruments might be addressed through research and then moves on to consider what specific questions should be included in the research programme of this thesis.

3.1 What form should a summative assessment process take?

3.1.1 General principles

An educational paradigm - "blueprinting"

To select appropriate assessment methods from a range of options many educational authorities (Bean, 1953; Hudson, 1973; Rowntree, 1977; Ward, 1980; Cangelosi, 1990) have advocated an approach that has come to be known as "blueprinting". In this approach a matrix is drawn up. One axis of the matrix consists of those attributes that are to be assessed. The second axis is made up of the range of possible assessment methods. For each attribute a decision is made about the most suitable method for assessing that attribute. This results in the development of an "assessment blueprint". This approach has two particular advantages. Firstly, it ensures that all attributes that are to be assessed are matched to an appropriate assessment method; conversely assessment holes can be identified - that is attributes for which no obvious assessment method currently exists and for which new methods need to be identified or developed. Secondly, it enables the identification of contextually-efficient assessment methods - that is methods which, in a particular setting, enable the assessment of a large number of the attributes under scrutiny. The rigour of this approach has been endorsed by those involved in the assessment of doctors (Hart, 1992).

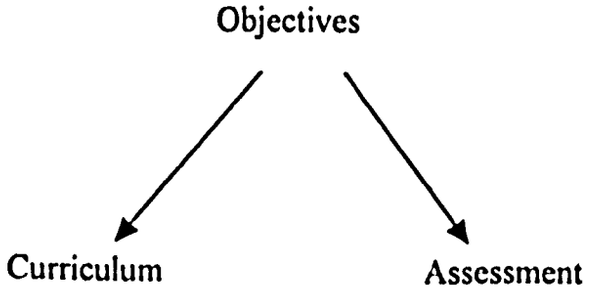
What this approach does require is both a comprehensive description of the attributes to be tested (the content of the assessment must be defined) and that suitable assessment methods are available or can be developed. Failure either to identify the content of the assessment or to provide suitable assessment methods would undermine the use of this approach.

Defining the content of the assessment

How might the content of a summative assessment process be defined? I believe that there are two approaches to this question.

The first approach, called here an "educational model", is to base the contents of the assessment process on the educational objectives of the training programme, in the same way that it might be expected that the curriculum of the training programme would be based on such objectives. This model has been termed "outcome-based" education (Spady, 1988); the term "educational model" is preferred because the second approach detailed below (p.41) is also a form of outcome-based education. The "educational model" is illustrated diagramatically in figure 3.1 below:

Figure 3.1. The relationship between educational objectives, curriculum and assessment in an educational model of defining the contents of a summative assessment process.



Such an approach has been advocated by Bloom et al. (Bloom et al. 1971) as a natural extension of the taxonomy of educational objectives originally described by Bloom and others (Bloom 1956; Bloom et al. 1964). These authors divide educational objectives into a "cognitive domain" (which includes knowledge and intellectual abilities/skills), an "affective domain" (which deals with the attitudes and values), and a "psychomotor" domain (that deals with technical (as opposed to intellectual) skills). The natural consequence of this approach is that an assessment process derived in this way will be centred on the assessment of these domains.

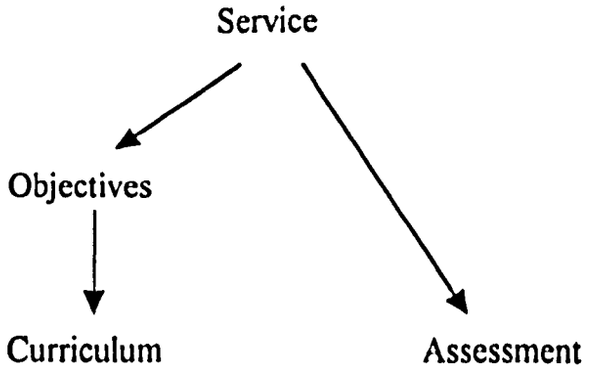
The strength of this "educational model" of defining the contents of an assessment process is that the educational intervention and the assessment process are inextricably linked through the educational objectives. Consequently, it might be expected that, provided the trainee follows the educational intervention diligently, there should be no problem in completing the assessment successfully.

There are two weaknesses to this approach to defining the content. The first is the risk that the assessment and the curriculum become directly linked - that is without reference to the educational objectives. Whilst this can be a strength (if there is a national form of assessment, all education institutions are likely to attempt to ensure that their curriculum will enable students to pass the assessment - the so-called "curriculum backwash" effect (Gipps and Stobart, 1993)) there is a real risk that the assessment itself becomes the prime driving force to the curriculum; "you get what you assess" (Resnick and Resnick, 1992). With time, if the requirements of the assessors change, the curriculum may come to have a more and more distant relationship to the original objectives.

The second problem occurs where education is vocational - that is, it is intended to equip a trainee to be fit for a particular purpose. Unless the educational objectives are clearly related to this purpose there is a risk that the assessment will fail to assure the fitness of the trainee for the purpose for which the training is designed; it will simply act as a measure of the effectiveness of the educational intervention in fulfilling its particular objectives. It is this issue that is central to the basis of the second model.

This second approach, called here the "vocational model", is to base the contents directly on the needs of the service that the trainee is entering. This is illustrated diagramatically in figure 3.2 below:

Figure 3.2. The relationship between educational objectives, curriculum and assessment in a vocational model of defining the contents of a summative assessment process.



In this model the crucial focus is the service for which the training is designed. It is this purpose that defines both the educational objectives (and thereby the curriculum) and the content of the assessment. Although there is still the potential risk of the assessment and the curriculum becoming directly linked, because the underlying link is less direct it is

likely that this risk is smaller - in this instance, if it is true that "you get what you assess" (Resnick and Resnick, 1992), the assessment is likely to drive the curriculum towards the needs of the service.

Such an approach is a derivative of the 'competence-based assessment' model advocated by Wolf (Wolf, 1995; Wolf, 1996). Originally advocated by Super as a "job analysis" approach (Super, 1949) Wolf defines competence-based assessment as "a form of assessment that is derived from the specification of a set of outcomes; that so clearly states both the outcomes - general and specific - that assessors, students and interested third parties can all make reasonably objective judgements with respect to student achievement or non-achievement of these outcomes; and that certifies student progress on the basis of demonstrated achievement of these outcomes. Assessments are not tied to time served in educational settings" (Wolf, 1995). I have used the title 'vocational' rather than 'competence-based' to signify the importance of basing the assessment on the specific requirements of the service for which the trainee is undertaking training rather than on a concept of competence that is not necessarily directly related to the needs of a particular service.

Some specific features of the vocational model merit highlighting. Firstly, the assessment process is less directly linked to the educational objectives; consequently this model may allow the assessment process to move closer to the strict definition of assessment of the individual rather than sitting closer to an evaluation of the educational intervention. Secondly, both the training and the assessment are driven by the needs of the service; as a consequence the control of the process lies less with the educators and more with the service providers and the consumers; as such it is likely that the validity of both the

education and the assessment will be improved. This second feature is of particular significance if the political and societal forces driving the revision of the system of regulating entry to general medical practice are to be addressed.

Derivations of this model could be applied at various stages in vocational training. In those professions in which there is some form of specialisation (e.g. the legal, medical, nursing and teaching professions) an "early vocational model" would be used in which the objectives and assessment were targeted at the requirements of initial entry to the profession (the assessment being designed to assess fitness to practice), whilst an "end-vocational model" would be used to set the objectives and assessment process for those entering specialised practice (the assessment being designed to assess fitness for purpose).

Selecting assessment methods

The second axis of the assessment blueprint matrix is the list of potential assessment methods. In selecting appropriate methods I believe that the categorisation of assessment methods into those which are primarily concerned with the assessment of "competence", and those which are primarily concerned with the assessment of "typical performance" is crucial. What is this difference, why does it matter, and what conclusions result from this distinction?

Messick defines competence as that which "refers to what a person knows and can do under ideal circumstances" whilst performance "refers to what is actually done under existing circumstances" (Messick, 1984). Wood and Power intertwine their definitions - "competence is the possession and development of sufficient skills, knowledge,

appropriate attitudes and experience for successful performance in life roles" - the implication being that competence is the "deep structure" that underpins the "superficial structure" of performance (Power and Wood, 1987). Although Miller suggests a further division - namely into competence, performance and action (Miller, 1990), with definitions of competence as the knowledge of how to do something, performance as the demonstration of how it is done, and action as the actual undertaking of that action, my view is that his definitions of competence and performance taken together are closer to the definitions of competence used by others, and that his definition of action is closer to the definitions of performance used by others.

Why does this distinction matter? The importance of making a distinction between the two types of assessment method lies principally in the practical issue of choosing the best method for the attribute requiring assessment - that is, to answer the question "is the attribute, or combination of attributes, best tested by considering competence or performance?". Tests designed to assess competence will consider the attributes without specific reference to the conditions in which the assessee will have to perform. Such tests can be designed to assess individual attributes that contribute to performance, for example aspects of knowledge, or of specific psychomotor skills (Anonymous, 1994). Consequently they may be particularly helpful in trying to diagnose where difficulties in performance lie. Furthermore, by designing such tests to examine very specific aspects of performance it is more likely that the influence of other aspects of ability on the performance in the test can be minimised thereby improving the reproducibility of the results in the test. The principal disadvantage of competence tests lies in their separation from reality - if tests are so separated from reality, how likely is it that they will predict performance in the usual setting with any degree of accuracy? As Cronbach states "all

decisions involve prediction; any test tells about some difference among peoples' performances at this moment; that fact would not be worth knowing if one could not then predict that these people will differ in some other performance or in the same performance at some other time" (Cronbach, 1964). Although the trend in test development seems to be to bring competence tests more close to the testing of performance - as Dwyer states (Dwyer, 1990) "We are clearly headed towards ..... assessing more complex samples of behaviour .... and making more realistic approximations to that actual behaviour during the testing process" - ultimately because such tests do not take into account the influence of the individual setting in which the assessee works on their performance such tests can only ever act as indicators of performance. Whilst competence tests may have a particular value in assessment processes that accompany training that is not targeted at a particular vocation (because they do not require the future context of performance to be defined), for the reasons outlined above their validity in vocational training settings is reduced.

Tests designed to assess typical performance consider the way in which the assessee performs within the environment in which they will be expected to work (Swanson et al. 1995). Such tests may offer a potential solution to the problem that arises from the difficulty in establishing the relationship between the results of a competence test and subsequent occupational performance (Vincent, 1996; Wolf, 1996). Further support for performance testing is offered by Gonczi who argues that this approach may help to overcome two particular risks associated with the assessment of complex activities (Gonczi, 1994): the first is the risk of reducing complex activities into multiple simple tasks that can be easily assessed in the examination setting but that rarely occur with such simplicity in the usual professional setting; the second is the risk associated with the

recognition that expertise is rarely generic but usually domain-specific - consequently the assessment of expertise can only be adequately assessed in the usual professional setting rather than in the examination setting. Work-based assessment may also offer one further advantage - namely a reduction in the dependence of assessment on the use of classical examination methods with, as a consequence, education and training being less driven by the needs of trainees to learn examination techniques (the "extrinsic rewards" effect noted on p.33). A report provided by the trainer is one such performance test.

The most convincing empirical support for the use of use of work-based performance tests has been provided by two pieces of work from the field of personnel psychology. In Asher and Sciarrino's review of 61 work sample tests (Asher and Sciarrino, 1974) they demonstrated that, of those factors considered, the best predictor of future job proficiency (measured using ratings by employment supervisors) was biographical information, followed closely by "motor work samples" (i.e. test of psychomotor skills as they would be performed in the usual work setting). Their results should not be a surprise. It seems intuitively correct that complex, multi-factorial information (such as biographical information and work sampling) is more likely to be predictive of future performance than single predictors (such as intelligence or manual dexterity). As a corollary of this, they argue that the greater the number of points in common between a test and the future activity (what they term as a "point-to-point theory"), the greater the likelihood of accurate prediction of performance. Their findings were strongly supported by the metaanalysis undertaken by Schmitt et al. (Schmitt et al. 1984). They demonstrated that "work samples....and supervisor/peer evaluations yield validities which are superior to those of general mental ability and special aptitude tests". These pieces of evidence provide strong support for performance tests in which the test mimics

closely the future employment setting, and thereby suggest that such tests may have particularly value when deciding whether or not an individual is fit for the purpose for which they are being trained.

The principal disadvantages of such tests are: that, because they examine overall performance (which may be influenced by a host of different attributes of the individual), it may be difficult to clarify exactly where a problem lies if performance is poor; that they require close observation of the assessee in the workplace (with the consequent risk of influencing the performance of the assessee); and that they may focus on what can be measured rather than on what matters (Eliot, 1991).

The above argument assumes that competence and performance are not linked in any predictable way, an argument originally put forward by Chomsky (Chomsky, 1965). However, I believe that there is a limited logical link between competence and performance. If, to paraphrase Messick, competence is "demonstrated action under examination conditions" and performance is "action in the usual professional setting" it is logical to argue that if an individual is able to demonstrate the correct action under examination conditions (i.e. is competent) that individual has at least some chance of undertaking that action in the usual professional setting (i.e. may be able to perform), whilst the individual who is unable to demonstrate the correct action under examination conditions (i.e. is incompetent) is unlikely to be capable of undertaking the action in the usual professional setting (i.e. is not able to perform) - although the possibility of an inability to demonstrate competence in an examination setting purely as a result of confounding factors (e.g. anxiety) must be borne in mind. Thus although competence may not directly predict performance, incompetence is likely to be a good (if not perfect)

indicator of poor performance - as Hart states "we should be ... surprised if it can be demonstrated that proven incompetence at the end of training turns into good performance in practice" (Hart, 1992).

In conclusion, whilst failure at tests of competence may predict poor performance, success in competence tests does not necessarily predict successful performance; success in performance tests is a better predictor of subsequent performance (Schmitt et al. 1984; Asher and Sciarrino, 1974). Consequently, particularly when the quest is the assessment of fitness for purpose at the completion of training for that purpose, whilst competence tests may have a place, particular emphasis should be placed on tests of performance.

The interpretation of the outcome of the assessment

Whilst the development of a blueprint ensures that all desired attributes will be assessed by an appropriate test, what approach should be used within these tests when making decisions about the individual's performance?

Two broad approaches to this problem are available (Cronbach, 1964). The first is simply a descriptive approach in which the results of the assessment are used to provide a description of the performance of that individual. The second is a comparative approach in which comparison is made with an agreed definition of acceptable levels of competence or performance. Because, as has been argued in chapter two (p.21), a principal function of a summative assessment process is the selection and deselection of individuals, a comparative approach will be needed.

Within such a comparative approach, Glaser has defined two possibilities (Glaser, 1963). The first is to compare the performance of the assessee relative to that of others - this is known as "peer-referencing"; the second, in which the performance is compared with some absolute standard of performance, is known as "criterion-referencing". A criterion-referenced test "is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (Glaser, 1971). If a summative assessment process is to be used in the selection of individuals, peer-referencing would be unacceptable both to those being assessed (because the selection of an individual will depend entirely on the performance of peers, an individual with consistent performance is more likely to be selected if the performance of peers is poor and less likely to be selected if the performance of peers is good, a system contrary to natural justice) and to those who will receive the service provided by those selected (the cut-off point above which individuals are selected will vary - there is no consistency to the standard of selection). To achieve the consistent standard of entry that is needed for a process that requires selection/deselection decisions to be made, criterion-referencing will be necessary (Livingston and Zieky, 1972). As Ebel states (Ebel, 1979), "criterion-referenced tests are best adapted to assist in categorical pass-fail decisions with respect to separate specific tasks".

3.1.2 The application of these principles within the setting of general medical practice in the United Kingdom

In this section the principles enunciated in the preceding section are examined with respect to their application in the specific setting of general medical practice in the U.K.

Definition of content for an assessment blueprint

It has been argued that, when the focus of assessment is one of "fitness for purpose", a "vocational model" (in which the content is defined by the needs of the service) should be the preferred model for the definition of the contents of the assessment process (p.42). Two arguments specific to general practice training support the use of such an approach.

Firstly, within the setting of medical training in general, such an approach is vigorously supported by Kane who argues that the contents of the assessment should be clearly related, either empirically or logically, to patient outcomes (Kane, 1982). Secondly, within the United Kingdom there is no nationally agreed set of educational objectives for training for general practice; it would not be possible to derive the contents of a national assessment process using the "educational model".

If a "vocational model" is appropriate, is it possible in this particular setting? Three published documents have relevance. At the time that the research component of this thesis was initiated only one description of the work of a general medical practitioner was available (although a further description was published in 1996 (The Nature of General Medical Practice Working Party, 1996)). This document, the Leeuwenhorst document (Statement by a working party of the second European conference on the teaching of general practice, 1977), describes a consensus view of the work of the general practitioner as agreed in the early 1970's; it is a brief and widely accepted consensus document, although it can be criticised for being very concise and thereby insufficiently specific particularly in terms of the clinical skills necessary for general practice. In addition two documents have been published which can be viewed as

attempting to define a curriculum for training based on the needs of the service. The Oxford Region Priority Objectives document (Oxford region course organisers and regional advisers group, 1985) was published in 1985; the focus of this document is an attempt to provide an overview of the aspects of general practice with which a doctor training for general practice should have become familiar by the end of vocational training. Although this document does go into considerably more detail about the skills needed for general practice than the Leeuwenhorst document, again it does not describe specific clinical skills that might be regarded as necessary for general practice. Conversely the Manchester Rating Scales (Centre for Primary Care Research, University of Manchester, 1988) is a list of skills (produced originally as a tool to be used by those training for general practice for self-assessment) which the authors consider to be important for a doctor to have mastered prior to entering independent practice.

Whilst the availability of these documents might provide a basis for defining the contents of an assessment process it is not possible to translate them directly into the contents axis of an assessment matrix. The reason for this is that, whilst there are areas of overlap between the three documents, none of the three documents include any weighting of the items. Kane argues that the aspects of performance that are to be assessed need to be weighted according to their relative importance for practice (Kane, 1982). In the particular context of a process that determines entry into independent general medical practice there would need to be a weighting of potential attributes focused specifically on their importance for independent practice. If these documents are to form the basis of an assessment blueprint based on the needs of the service further work would be necessary to clarify which areas of content are of greatest importance for independent general medical practice.

Selecting assessment methods

In the setting of general medical practice, what options for assessment methods might be available for selection?

Within medical practice as a whole there is now a considerable repertoire of assessment methods (Anonymous, 1994). All are tests of competence, although in a number of instances attempts have been made to make them much closer to tests of performance (through the use of, for example, simulated patients (Rethans et al. 1991)).

Within general medical practice Rethans has provided evidence that suggests that performance tests may be particularly important (Rethans et al. 1991). Using simulated patients inserted within the normal consulting timetable for general practitioners, he demonstrated that there was no direct relationship between the performance of doctors and their competence as demonstrated by classical competence tests. This evidence, provides empirical evidence to support the hypothesis that performance tests may have a particularly strong role in this particular setting (p.45).

As part of the original proposals for summative assessment, the JCPTGP suggested consideration of four particular assessment methods (Joint Committee on Postgraduate Training for General Practice Working Party on Assessment, 1992a): a multiple-choice questionnaire to assess knowledge and its use; an assessment of performance in the consultation using either video-taped consultations or simulated patients; a written submission based on medical audit; and a report provided by the trainer. Of these, two

(questionnaire and written submission) are competence tests and two (trainer's report and consultation assessment) are performance tests.

The possibility of including a trainer's report as a component of a summative assessment process in general practice has been strongly supported by a number of authorities who have come to recognise the pivotal role that a trainer could have in assessment in general practice (Harden, 1979; Kenyan et al. 1987; Newble, 1988; Carline et al. 1989; Carney, 1992; Difford and Hughes, 1992). There seem to be two main reasons for this. The first is a recognition that the long-term relationship between the trainer and trainee in general practice provides an opportunity to include an assessment based on performance throughout the training year, rather than assessment based on a single-point test (Difford and Hughes, 1992). The second is the recognition that, whilst some tests would only allow assessment of highly selected aspects of performance, a report provided by the trainer could potentially provide an assessment based on a large number of aspects of the work undertaken by a general practitioner (Kenyan et al. 1987) i.e. it is contextually-efficient (p.38). Although doubts do exist about the weight that should be attached to an assessment undertaken by an individual general practitioner trainer (Miller, 1990) and little published evidence on the use of such a report exists (Campbell et al. 1993) it is apparent that the possibility of including a trainer's report within the assessment blueprint is worthy of consideration.

The interpretation of the outcome of the assessment

The only experience available within general practice in the U.K. of a referenced assessment is the membership examination of the Royal College of General Practitioners. Unfortunately this examination is peer-referenced (Haslam, 1998), an approach which

has been argued (p.48-9) as being unacceptable for a summative assessment process. If a summative assessment process is to be adopted, absolute standards will need to be set.

In the selection or development of suitable standards, at what level should such standards be set and who should set them? If the prime purpose is to select out those who are not yet ready for independent general practice, the logical answer is to set the standards at the minimum for such practice. Standards might be set by experienced members of the profession (including teachers within the profession), by the public (who are to be protected) or by the assessees themselves. It has been argued (p.29) that, with adequate safeguards, there remains an argument in favour of self-regulation by the professions. If this argument is accepted it would be reasonable for members of the profession to set the standards but consideration will need to be given to whether or not it is possible to include the views of the public in the setting of such standards.

It is important to clarify that in using the word "minimum" this is not the same as "minimal". This is not just a semantic issue. Every assessment process that results in selection (or deselection) will need to have a minimum standard; if this is associated with criterion-referencing the standard can be defined as the absolute minimum level below which a person wishing to enter independent general medical practice cannot fall. This is not the same as having minimal standards; the adjective minimal in this context implies movement towards the lowest possible standards. If the word minimal is used to replace minimum, it is possible that summative assessment would be perceived as a process the intent of which was to drive down standards rather than to maintain standards - the minimum standards could be perceived as the maximum standards. For this reason, I believe that it is important to retain the word minimum when discussing standards. Such

a description does not preclude the possibility, over time, of driving the minimum standard upwards, an approach which might then be legitimately described as optimal.

3.1.3 Conclusions

An assessment blueprint offers a structured approach to the development of a schedule for a summative assessment process. If a blueprint is to be developed, the attributes to be assessed will need to be defined and suitable assessment methods selected or developed. A method for interpreting the outcome of the assessment is then required.

When assessment of fitness for purpose is the desired goal, it has been argued that the content should reflect the needs of the service, tests that enable the assessment of performance take on an increasing importance, and the interpretation should be based on comparison with absolute standards (criterion-referencing).

Within the specific context of assessment at the entry to independent general medical practice in the U.K., from the descriptions of potential content that are available selection of contents based on their importance for independent general medical practice will need to be made and matched to suitable assessment methods. There is evidence that performance tests are of particular relevance in this setting, and there is considerable support for the adoption of a trainer's report as a performance test. Criterion-based standards will need to be set at the level of the minimum standard for independent practice.

If a trainer's report is worthy of further consideration as an assessment method, what technical requirements would need to be met by such a report?

3.2 What are the technical requirements of assessment instruments?

In his classic review of psychological tests Cronbach delineates the technical requirements of such tests (Cronbach, 1964). Similarly, Fabb and Marshall have considered in detail the requirements of tests designed specifically for the testing of doctors in the setting of general practice (Fabb and Marshall, 1983). In broad terms the requirements listed in these two publications fall into two categories, the categories being based on two groups of individuals who might be considered to have a legitimate concern. The first category consists of those requirements which focus on academic rigour; the second category consists of those requirements which focus on the utility of the test.

3.2.1 Issues of academic rigour

Most authorities on assessment recognise two major components to the academic rigour of a test - validity and reliability (Bean, 1953; Chauncey and Dobbin, 1963; Cronbach, 1964; Hudson, 1973; Ebel, 1979; Ward, 1980; Fabb and Marshall, 1983; Ward et al. 1996). Validity is concerned with the extent to which the test measures the true position; reliability is concerned with the extent to which the test produces reproducible results. Although the two are usually separated, it must be borne in mind that an unreliable test can not be a valid test - a test that fails to measure some attribute with any level of consistency can not be said to be measuring anything at all (Stanley and Al-Shehri, 1993).

In the context of assessment instruments, validity has usually been broken down into the four categories originally described by the American Psychological Association (content,

construct, predictive and concurrent) (Ward et al. 1996). Content validity is the extent to which the instrument tests the attributes that are important. Construct validity is the extent to which the result in the test is linked to the underlying psychological phenomena which it is intended to measure. Concurrent (sometimes known as criterion-related) validity is the extent to which the result in the test concurs with other measurements of the same attribute. Predictive validity is the extent to which the result predicts performance in the future.

In addition, three other categories of validity are in common use in this context. Although an elemental approach to validity seems to be well established, a number of authors (Cronbach, 1964; Ebel, 1979; Crocker and Algina, 1986; Thorndike, 1997) also highlight a broader issue which, in this thesis, will be termed "overall validity". They argue that the crucial issue to address is probably not "whether this is a valid test", but "how valid is this test for the decision I want to make?"; that is, the overall validity of a test concerns the extent to which the test enables the required decision to be made. Face validity is a term that is used to describe whether or not the test appears to be reasonable (Fabb and Marshall, 1983; Streiner and Norman, 1995); it is often used when a test is being considered by those who are likely to be using it. On balance I believe that this is very similar to overall validity, but that the term overall validity is the preferred nomenclature as there is less implicit dependence on what the test looks like and more dependence on how it performs. A further form of validity has been described more recently, particularly in the setting of medical education, which is concerned with the effect of the test on the curriculum (Newble, 1988). This is the extent to which the assessment process affects the curriculum of associated training programmes.

Reliability has also been subdivided into categories. These are stability and equivalence (Glaser, 1963; Hudson, 1973; Fabb and Marshall, 1983; Streiner and Norman, 1995). Stability is the extent to which the results of the test remain stable when applied in different settings; this concept includes the issues of intra-rater reliability (same assessor, different settings) and inter-rater reliability (same setting, different assessors) (Mulholland and Tombleson, 1990); it also incorporates the variability in performance of the individual being assessed (Hudson, 1973). Equivalence is the extent to which there is agreement on the result for items that seem to be measuring (or that are designed to measure) the same attribute; when applied within a single test format, it is known as internal consistency; when applied between different tests, it is the same as concurrent validity.

One other issue of academic rigour has been suggested by Fabb and Marshall - namely the degree of objectivity of the test (Fabb and Marshall, 1983). The main rationale proposed for measuring objectivity is that objectivity is associated with reliability; as De Groot argues "by ensuring that the assessor behaves in such a way as to preclude interference of the personal opinions, preferences, modes of observation, views, interest, or sentiment" it is more likely that the results will be reproducible (De Groot, 1969). However, this view has been challenged (Van der Vleuten et al. 1991), the authors of that challenge concluding that "the particular assessment method chosen, whether principally objective or subjective, should depend on the particular testing situation". Extending this argument, if a test is reliable whether it is subjective or objective in method is not relevant - objectivity is subjugate to reliability. Consequently, I do not believe that objectivity should be included as an additional requirement to that of reliability.

3.2.2 Issues of utility

The second group of requirements centres on whether or not an assessment method can actually be used by those for whom it is intended. Cronbach suggests two tests of utility (Cronbach, 1964). Applicability is the ease with which the method can be applied; it includes the feasibility of using the method (that is, what is possible and practical as opposed to the ideal (Dauphinee et al. 1994)), and the acceptability of it to those for whom it is intended. Efficiency is the cost (in particular in time and finances) of applying the instrument.

3.2.3 Relative importance of the requirements

A number of requirements have been listed. In this section an analysis is made of the relative importance of these requirements with particular reference to the development and testing of a performance test such as a trainer's report.

Content validity:

It has already been argued that, in developing an assessment blueprint, a definition of the content of the assessment process is crucial. A decision is then made as to which attributes are to be tested by which assessment instrument. Without a definition of the content area(s) to be addressed, the instrument will be meaningless. If a rigorous method for the selection of contents has been made in the development of an instrument, a formal test of the validity of the contents may not be needed. However, in two circumstances such a test may be necessary. Firstly, if there are any concerns about the rigour of the method used for the selection of contents, a specific test of content validity will be needed. Secondly, Popham and Husek (Popham and Husek, 1969) argue that whilst the

emphasis in a peer-referenced instrument is on maximising variability of results for each of the items within the instrument (thereby enabling differentiation of candidates), for a criterion-referenced instrument variability of results is less important (because, by definition, the aim is simply to measure the attribute against a predetermined standard) but that there should be a much greater emphasis on ensuring that the correct domain of content is being assessed (to ensure that the correct areas are being assessed).

Overall validity:

In the development of an instrument it is crucial that attention is given to ensuring that it is likely that the instrument will enable the proposed assessment to take place. In developing a performance test it is logical to predict that this is most likely to happen if the conditions in which the test is undertaken mimic closely the conditions under which the assessee would usually be expected to perform; when testing the instrument, it would be crucial to be able to demonstrate that the test will actually do what it is supposed to do.

Construct validity:

For any psychological measurement Cronbach and Meehl (Cronbach and Meehl, 1955) argue that it is important to establish that the measurements made do actually measure the psychological phenomena (i.e. the constructs) that underpin the attributes measured. Ebel counters this by arguing that construct validity is of little importance in performance tests - if the performance is demonstrably happening, does it matter whether or not the postulated underlying attribute actually exists? (Ebel, 1979). My view is that Ebel's argument is the more convincing in the context of instruments that are specifically

designed to test performance and should therefore prevail i.e. that overall validity (the outcome measure) is more important that construct validity (the process measure).

Predictive validity:

For an assessment process the issue of predictive validity is crucial (Bloom, 1968). In the development of an instrument it is likely that predictive validity will itself be dependent on content validity and construct validity - an instrument that covers the attributes that are important (i.e. has content validity) and which concentrates on the psychological phenomena that determine continuation of that behaviour (i.e. examines a construct that determines continuing behaviour) is likely to predict future performance. There is, however, strong evidence that future performance is predicted most effectively by previous performance (p.46; Schmitt et al. 1984; Asher and Sciarrino, 1974); as a consequence an assessment method that enables accurate measurement of current performance (i.e. a performance test that has overall validity) should offer predictive validity. It has already been argued above that, for a performance test, construct validity is subjugate to overall validity. Consequently, my view is that in the development of a performance test overall and content validities should be the primary goals in the development of the instrument. Because of the importance of prediction consideration must be given to a formal review of predictive validity within the testing of the instrument.

Concurrent validity:

Whilst comparison with existing instruments may be of some help in the design of an instrument, concurrent validity is probably more important in the testing phase. However such a measurement can only be undertaken if there exists already an instrument that has

been shown, or is widely accepted, to be valid. If no such test exists, then a test of this form of validity is of lesser importance.

Curriculum effects:

It has been argued that, when fitness for purpose is the desired goal of an assessment process, the selection of content should be based on a vocational model and performance tests are of particular importance (p.55). In the vocational model for selecting contents, the link between assessment and the curriculum is indirect, through a common focus on the requirements of the service (p.41). If both the assessment process and the curriculum are based on the needs of the service (i.e. have content validity) they should be related. Consequently in the development of an assessment process through a vocational model, curriculum effects are subjugate to content validity. Once content has been selected appropriate instruments will need to be selected. Because the tests themselves may have an effect on the curriculum (p.34) consideration should be given to a test of the degree to which the instruments themselves have effects on the curriculum.

Stability:

The balance between the importance of inter-rater and intra-rater reliability will depend on the purpose of the assessment instrument. An instrument designed for the testing of multiple individuals by a single assessor (for example, a national written test) must have intra-rater reliability. For an instrument designed for the testing of an individual by one or more assessors, the issue of inter-rater reliability becomes much more important - the assessee needs to be sure that (s)he has an equal chance of success irrespective of the assessor. Furthermore, inter-rater reliability is likely to be dependent on adequate intra-rater reliability - it is unlikely to be possible to achieve inter-rater reliability if there is no

intra-rater reliability (if the instrument encourages inconsistency within assessors there is unlikely to be any consistency between them - two wavering opinions are unlikely to waver consistently); conversely, it may be possible to have intra-rater reliability without inter-rater reliability (assessors may be consistent in their own views, but their views may be a consistent distance apart). Consequently, I would argue that for all assessment instruments, but particularly for those which will be applied by large numbers of assessors, the most important issue is inter-rater reliability - that is, the likelihood that a trainee will stand an equal chance of passing the test irrespective of the assessor completing it. Attempts should be made to promote inter-rater reliability within the development of the instrument but, because of its importance, a test of the inter-rater reliability of the instrument will be an important component of the testing phase. The importance of assessee variability is also likely to depend on the nature of the instrument; instruments that consider performance on a single occasion are likely to be much more vulnerable to this effect than performance tests that are based on a conclusion that draws on observation of several examples of performance.

Equivalence:

Internal equivalence (i.e. within the instrument) only arises when it is intended to assess the same attribute on more than one occasion within the instrument. External equivalence (i.e. with another instrument) within a summative assessment process might be considered to take two forms - comparison with other instruments within the same process, and comparison with instruments that are not part of the process. The latter of these is essentially an issue of concurrent validity. The former issue deserves consideration - to what extent is overlap between different components of an assessment process desirable? If it is desirable, then overlap should be sought and demonstrated. I

would argue that there should be overlap when two conditions are met: firstly, that it is an attribute of high importance; secondly, that either both assessment instruments have been demonstrated to have similar levels of validity and reliability, or where these properties are unknown for both instruments. If the attribute is of relatively low importance, or if one method is demonstrably better, there should be no overlap.

Applicability:

It is important that the instrument can be applied in the setting for which it is intended, and that it is likely to be acceptable to those involved. If these issues are not addressed, then however academically strong the instrument is, it is likely to prove very difficult to get it widely adopted. Acceptability is likely to be dependent not only on whether the test is perceived, or measured, to be valid and reliable, but also on whether it is feasible in use. It is therefore important to ensure that feasibility has been demonstrated.

Efficiency:

Whilst any instrument should be efficient, it can be argued that this is one component of applicability (Norman et al. 1985) - a test that is very consuming of financial, personnel or time resources is not likely to be feasible, and thereby not acceptable. Because efficiency can be considered as one element within feasibility I would argue that a test of efficiency additional to one of feasibility is not necessary.

3.2.4 Conclusions

It is concluded that for performance tests the following technical requirements are of particular importance: overall validity, content validity, predictive validity, curriculum effects, inter-rater reliability, and applicability/feasibility.

Of these requirements some have particular importance in the development of an instrument (overall validity, content validity, inter-rater reliability and applicability/feasibility) and some in the testing of the instrument (overall validity, content validity, predictive validity, curriculum effects, inter-rater reliability and feasibility).

Some requirements are only applicable if certain conditions are met: concurrent validity would only be included within the testing phase if a test of demonstrable validity was available; internal equivalence would only be included in both phases where overlap was felt to be desirable.

3.3 Current experience of reports provided by trainers

Having examined the technical requirements of assessment instruments with particular reference to performance tests, this section now focuses in particular on the place of a trainer's report as one such performance test through an examination of experience gained with such reports as assessment instruments. In particular, it considers the extent to which the technical requirements listed above have been met.

3.3.1 Experience outside general medical practice

A literature search undertaken in 1994, and repeated at the end of 1997 (appendix 1.1), revealed evidence of trainer/supervisor assessment in a number of professional settings - social work (Akhurst, 1978), radiography (Graham, 1981), nursing and midwifery (Phillips, 1993), management (Barker, 1993), business work experience (King and Danks, 1986), industrial work experience (Rakowski, 1990) and teaching (Preece,

1993). Of these, in only two instances is there published evidence of a structured report provided by the trainer/supervisor being used for the purposes of summative assessment of the trainee (Rakowski, 1990; Preece, 1993).

Rakowski reports the experience from Brunel University in which those supervising the industrial work experience of trainees in engineering were asked to complete a structured report (Rakowski, 1990). The views of the supervisors were sought in twelve areas - namely ability to carry out assigned tasks, technical competence, problem-solving ability, willingness to accept responsibility, initiative, self-organisation, perseverance, capacity to innovate, interpersonal skills, quality of oral reporting, quality of written reporting and quality of the logbook kept during the placement. Supervisors were asked to indicate their assessment using a five-point scale. The results of this report contributed approximately 7% to the final marks assigned in the examination. An analysis of this report against the technical requirements listed above demonstrates that there are problems with this report: the basis of the contents of the test is not made explicit, and no standards on which to base the marks given is provided. Whilst there is no evidence of any formal testing of the report form to see if it actually works the evidence that is available suggests that the report form was used which suggests that it was, at least to some extent, feasible.

Preece describes a very similar activity in the context of the teaching practice performance of students completing their teacher training (Preece, 1993). He uses the Exeter Teaching Practice Schedule, a structured report in which the supervisor (a university tutor) is asked to rate the student on twelve aspects of performance falling into three broad categories. Each of these aspects is accompanied by descriptions of three

stages of progression along a dimension (fail, weak pass, strong pass) that can be used by the supervisor in reaching their assessment. Testing of this report form has demonstrated that it has overall validity, concurrent validity (the concurrent measure being the views of experienced teachers) internal consistency and is feasible. There are again some problems with this report: no evidence is provided on the way in which these standards have been drawn up and the basis of the contents is not made explicit (although it is implied that it is based on the views of the supervisors as to the requirements for a trainee entering the teaching profession).

3.3.2 Experience within general medical practice

At the beginning of this project, in 1993, two full draft trainer's reports were already in existence - namely the pilot assessment package from the west of Scotland (Campbell et al. 1993), and the North Thames (West) regional trainer's report (Rhodes and Styles, 1995) - and a number of other groups were beginning to develop potential report forms (e.g. the Trent Regional trainer's report and the Reading trainer's report (Hasler, 1994)).

The West of Scotland package

The west of Scotland package contains a report by the trainer (Campbell et al. 1993). However there are concerns about the overall validity of this report because of the rarity of failures in this component of the package (Campbell and Murray, 1996). Although it is difficult to be certain of the reasons for this there are two apparent areas of concern. Firstly, in this report form the trainer is simply asked to assess the trainee in five broad areas of performance; it may be that trainers are uncertain as to exactly what should be included in the assessment in each of these areas - that is, there is an issue about the content validity of the instrument. Secondly, trainers are simply asked to rate whether or

not the performance in that area is satisfactory without any clear indication of the standards against which the trainee should be assessed.

The North Thames (West) trainer's report

This development is based on the work of a regional group who invited trainers to set the contents of the report (Rhodes and Styles, 1995). For each item standards were set in the form of "exemplars" of what would constitute acceptable performance and what would constitute unacceptable performance. The major strength of this report is its emphasis on the trainer undertaking observations of typical performance - it is specifically designed as a test of performance observed in the usual professional setting rather than as a competence test. It has already been argued that this is likely to enhance its overall validity. Nevertheless, there are a number of problems with this report. Firstly, because this report was developed in a single region of the United Kingdom it is conceivable that the contents have a particular local flavour. Secondly, the use of exemplars of performance only at either end of a scale (i.e. definitely poor and definitely good) results in a risk that the assessor (i.e. the trainer) is still left with the major problem of trying to come to a decision when the performance of the trainee lies at, or very close to, the borderline between pass and fail; this leaves considerable room for judgement, which might be expected to affect inter-rater reliability. Thirdly, and most importantly, this report form has been subjected to no formal testing; it is not possible to describe how it performs when it is used.

Other projects

An informal survey of Directors of Postgraduate General Practice Education undertaken in October 1993 into aspects of summative assessment enquired about the development

of trainer's reports (Hasler, 1994). Fourteen of the 28 Directors replied. Four regions were developing systematic trainer's reports, but all were at a rudimentary stage of development and were untested.

3.3.3 Conclusions

Experience with trainer's reports as an assessment instrument does exist, but there is limited published evidence about this experience. Evidence from outside of medicine demonstrates that it is possible to design trainer's reports that are feasible, have overall validity, concurrent validity and internal consistency; the evidence on the validity of the contents is questionable, and neither predictive validity nor inter-rater reliability have been formally tested (Rakowski, 1990; Preece, 1993). The evidence from within general practice is also limited (Campbell et al. 1993; Rhodes and Styles, 1995); the only major study in this area (Rhodes and Styles, 1995) describes the selection of content, a process that is likely to enhance the overall validity of the report, and a set of standards for a trainer's report. However, there are considerable limitations to the methods used for the selection of contents, there are concerns about the feasibility of using the standards provided for assessment purposes, and no tests of the properties of this instrument have been made. It is concluded that suitable trainer's reports can be developed but that if a trainer's report is to be used as a component of a national summative assessment process for the selection of doctors for independent general practice, the major limitations of the instruments currently available mean that they should not be accepted as they stand.

This leaves two options. The first is to build on the current report forms, the second is to develop and test a new trainer's report. Whilst the former approach has the advantage that some work would not be needed, because there appear to be fundamental problems

with the design of the proposed reports (in particular the validity of the contents and the feasibility of the proposed standards) there is a risk that further work based on this foundation will still not enable the requirements for assessment instruments to be met. It is therefore suggested that a new trainer's report is needed. If such a report is to be developed *de novo* and tested, what questions will need to be answered? In the final section of this chapter, this issue is examined.

3.4 What research questions need to be answered in the development and testing of a new trainer's report for summative assessment in general practice?

It has been argued in the previous section that a new trainer's report is needed for summative assessment in general practice. This final section considers the research programme that is needed to provide such a report. The section begins with a broad question and, through a consideration of the issues that have been raised in the earlier parts of this chapter with particular reference to general practice in the United Kingdom, develops a specific and answerable hypothesis. The section concludes with an overview of a programme of research studies to test this hypothesis. This research programme then forms the research component of this thesis (chapters four and five).

3.4.1 The broad question

If a new trainer's report is needed the most important question that needs to be answered is "can a report form be designed that adequately fulfills the technical requirements of assessment instruments?". In the context of summative assessment for general practice in the U.K., whilst this broad question may be laudable, is it actually answerable?

3.4.2 Selecting answerable questions

It has already been argued in the preceding section that some technical requirements are of particular importance, some in the development of the instrument and others in the strategy for testing that instrument (p.65). Having considered these issues in general terms, what form do these issues take in the context of a trainer's report as part of a summative assessment process for entry to independent general practice?

Development phase:

Content validity:

The selection of the contents of the assessment process is vital. It has already been argued that the selection of contents has two components. Firstly, because training for independent general practice is vocational, the contents of the whole summative assessment process will need to reflect those attributes judged to be most important for the service - that is, for independent general practice. Secondly, there will need to be a process of blueprinting - that is, a process of matching the selected contents for testing with suitable assessment instruments. The instruments suggested by the JCPTGP (Joint Committee on Postgraduate Training for General Practice Working Party on Assessment, 1992a) offer a starting point for this selection. Within that blueprinting, it has been argued that, if some attributes are both deemed to be of extremely high importance and might be equally assessible by more that one method, there may be a place for overlap between instruments.

Overall validity:

For summative assessment in general practice the term overall validity is concerned with whether or not the process, and a trainer's report in particular, will actually measure what is required - namely the separation out of those doctors who are not yet ready for independent general practice. It has been argued that future performance is most likely to be predicted when the conditions in which the assessment is undertaken mimic closely the usual professional setting (p.60) - that is that the assessment is based on evidence that is truly representative of the usual performance of the trainee. Whilst a trainer's report may look appealing in its likelihood of fulfilling this requirement it will be important that, within the development phase, consideration is given to ensuring that the report form does require observation of performance and that absolute standards are provided against which performance can be compared (these standards being based at the level of the minimum for independent general practice) (p.54).

Inter-rater reliability:

For an instrument that would be applied by up to 3500 trainers it is crucial that the trainees can be sure that they stand an equal chance of passing irrespective of who the trainer is - that is, it must have inter-rater reliability. In chapter two (p.20) it was argued that there are two crucial components to any assessment; the first is the gathering of evidence, the second is the interpretation of that evidence. Consequently, it is likely that inter-rater reliability will be enhanced if, in the development phase, consistency in the gathering of evidence and in the making of judgements is encouraged.

Applicability:

In the development of a trainer's report, feasibility and acceptability are likely to be enhanced if there is a process of consultation between the instrument designers and those who will be using the report - in this instance the general practitioner trainers and trainees (Rhodes and Styles, 1995; Rhodes, 1998). Feasibility is also likely to be enhanced if all trainers using the report form understand exactly how it is to be used - there will need to be clear guidance.

Testing phase:

Predictive validity:

Whilst this is probably the most important form of validity of any assessment instrument, there are three practical issues that mean that such a test may prove impossible, at this stage, in the context of a trainer's report for summative assessment in general practice. Firstly, anyone who fails summative assessment would not be able to enter independent general practice; it would not be possible to reanalyse at a later date the performance of those who fail. Secondly, in the U.K. there is currently no performance-based continuing assessment process for established general practitioners against which the results of summative assessment could be subsequently compared; until such a system is in operation it will not be possible to review directly the results of the components of summative assessment for evidence of predictive validity. Thirdly, to obtain a useful indicator of predictive validity it would be necessary to consider performance some time after completion of summative assessment (for example, at five to ten years), a timescale beyond the scope of this study. Difficulties in the measurement of predictive validity are well recognised (Wolf, 1996) and, unfortunately, it may have to be accepted that such a test is not possible within the research proposed.

Overall validity:

To demonstrate whether or not the instrument has overall validity it will be necessary to demonstrate that this instrument does allow the separation out of a group of doctors - that is, it must have discriminatory power. To demonstrate this it will be necessary for the report form to be tested on a group of trainees to see if trainers would recognise any of them as being unready for independent general practice.

Inter-rater reliability:

The major difficulty in undertaking a measurement of inter-rater reliability for a trainer's report in general practice is that each trainee only has one trainer at a time. Consequently, if inter-rater reliability is to be measured, ways of comparing the assessments made by more that one trainer on the same trainee which do not disturb the normal training relationship and environment will need to be found.

Feasibility:

This is the extent to which it is possible to use the instrument under the conditions for which it was intended. This could only be done by testing the form under such conditions and seeking information on the strengths and weaknesses of the form when used in these conditions.

Curriculum effects:

The absence of a uniform curriculum for general practice training would make it difficult to test the degree to which the current curriculum has been disturbed by the introduction of a summative assessment process. Similarly, although it should be possible to

demonstrate whether the introduction of such a process does required the introduction of new components to the curriculum to enable the technical issues of the process to be addressed, this effect could only be accurately measured after the introduction of the process on a national basis. Such a study could consequently only be undertaken after national adoption of the process, a timescale beyond the scope of this study.

Content validity:

It has been argued that two factors would influence the presence of a formal test of content validity - the rigour of the method for initial selection of the content of the report form, and the particular importance of content in a criterion-referenced instrument.

Other issues:

Concurrent validity:

Within general practice in the U.K. there is no current assessment process testing at the level of minimum standards. Whilst there is the examination for the membership of the Royal College of General Practitioners, this is a test specifically set to consider optimum standards (Haslam, 1998), and is a test of competence. The only other alternative is to consider the possibility of testing consistency with other instruments within the summative assessment process. Whether or not this is possible depends entirely on the assessment blueprint. If different assessment instruments are deliberately targeted at the assessment of different attributes concurrent validity would not be expected; if overlap is intended, then a measurement of concurrent validity is reasonable. It will only become apparent as to whether or not a test of concurrent validity is possible once the content issue has been addressed. This issue is reconsidered in chapter six.

Internal consistency:

There is a separate issue of consistency within the trainer's report. Again, this is entirely dependent on the extent to which there is overlap, this time within the report form itself. This issue is also reconsidered in chapter six.

3.4.3 Conclusions and a research hypothesis

From this analysis of the effect of the particular context on the development and testing strategies, it is apparent that, with the exception of predictive validity and curriculum effects, all the technical requirements selected as being important for new assessment instruments earlier in the chapter could be addressed. Consequently, my view is that the following constitutes the desirable, and potentially answerable, hypothesis for the research programme:

Is it possible to develop a report form, to be used as part of a process of regulating entry to independent general medical practice in the United Kingdom, that will enable aspects of the knowledge, skills, attitudes and practice of the general practitioner trainee to be assessed by the general practitioner trainer in a way that has demonstrable levels of important aspects of feasibility, reliability and validity?

When compared with the literature outlined above (p.66-9) I believe that the unique contribution of the proposed research lies in the attempt to develop and test as comprehensively as possible a report form for use by the trainer in the particular setting of medical education. It is also hoped that the research will augment the evidence currently available and thereby provide information that is more widely applicable to the role of trainer's reports as part of assessment processes.

3.4.4 A programme of research

To answer this hypothesis a programme of linked research studies is proposed. These are:

I. Development of a trainer's report:

A study to determine an appropriate structure for the trainer's report - this study would consider the views of trainers on the requirements for the structure of a new trainer's report; the results of this study should enable a report form to be designed that promotes feasible assessment (the premise being that by incorporating the views of trainers feasibility is likely to be enhanced (p.73)). The results of this study, in conjunction with theory and experience from elsewhere should enable the development of a report form that promotes overall validity by encouraging assessment through the observation of performance. It is also intended that this process of consultation with trainers should be used to ensure that the rest of the research programme is likely to answer the most important issues - that is, it will enable the validity of the research proposals to be tested.

A study to select the contents of a trainer's report - this study would enable appropriate content for a trainer's report to be selected as part of a summative assessment blueprint. In this study the concept of a service-based assessment blueprint (the "vocational model" (p.41)) would be used. Attributes required by practitioners in order to deliver the service would be selected on the basis of the views of one group of key players (general practitioner trainers) of their importance for independent general practice. Opinion would also be sought on appropriate instruments for the assessment of those attributes.

A study to assess the content validity of the report form - because of the importance of the issue of content it is proposed that, as part of the development phase, a formal test of content validity is undertaken. This study would assess the views of another group of key players on the validity of the content of the proposed trainer's report. This proposal is particularly included because of the risk of introducing bias through the exclusive use of one group of key players in the selection of content.

A study to develop appropriate standards for use in the report - a criterion-referenced assessment instrument will require the setting of standards set at the level of minimum standards for entry to independent general practice. This study would enable the development of minimum standards for each of the attributes selected for inclusion in a trainer's report. It would be designed to enable those involved in training to set standards at this level, and would also consider the ways in which testing against these standards could be realistically undertaken.

The combined results of these four studies should enable the production of a draft trainer's report suitable for testing.

2. Testing of the report form

A study to assess the overall validity (discriminatory power), feasibility and inter-rater reliability of the report form - this would take the form of a field test. It would assess whether or not the instrument does enable the separation out of any doctors (as a measure of the overall validity of the report form), and whether or not the report form is feasible in use. In addition, by enabling more than one trainer to undertake an

assessment of a trainee at the same stage of training, the degree to which raters using the report form agree on the result of the assessment would be measured.

It is this programme that forms the basis of the research component of this thesis in chapters four and five. Chapter four analyses methodological issues and describes the methods chosen for these studies. Chapter five details the results of the studies. Chapter six discusses the findings from these studies and examines them in the light of other work, whilst chapter seven returns to the wider issues in relation to a trainer's report and summative assessment.

CHAPTER FOUR - METHODS

In the concluding section of chapter three a series of linked research studies have been proposed. This chapter is concerned with the methods by which those studies will be undertaken. The chapter is divided into two sections. The first section examines methodological options within the context of the development and testing of a new trainer's report. In the second section, an analysis of the issues that arise in the particular setting of training for general medical practice in the U.K. is made to enable the selection of suitable methods to fulfill the aims of the proposed studies; the methods selected for each study are then described in detail.

4.1 General methodological themes

The programme of research studies delineated at the end of chapter three consist of two main types of study - those which are primarily concerned with seeking the views of those who might be considered to have a major stake in the place of a report provided by the trainer and those which are concerned with measuring properties of the report form in use. This categorisation of the types of research study is shown in table 4.1 (overleaf).

Table 4.1: the types of study contained in the research proposals.

Studies in which views are sought	**Studies in which properties are tested**
Determining an appropriate report structure	Test of content validity
Selection of contents and assessment methods	Test of inter-rater reliability
Views on content validity	Test of overall validity (discriminatory power)
Development of standards	
Views on feasibility	

4.1.1 Seeking views

Views can be sought in a number of ways (Streiner and Norman, 1995). These methods divide into those in which views are sought directly (typically by face-to-face or telephone interviews), and those in which views are sought indirectly (typically by the administration of questionnaires). The relative strengths and weaknesses of these methods have been considered in detail by Streiner and Norman (Streiner and Norman, 1995). Their conclusions, along with additional evidence of particular importance in the setting of the development and testing of assessment instruments, are summarised below.

Interviews:

Interviews encourage responses to all questions - it is difficult to ignore questions when talking with an interviewer (Quine, 1985). The interviewer can clarify questions if necessary and can also clarify responses by further questioning; interviewers can also help interviewees pick their way through complicated sequences of questions.

Conversely, interviews are costly in terms of interviewer training (to ensure consistency in questioning and recording) and interviewer time. Interviews also require time for organisation both to ensure that the respondent will be available and to minimise the risk of the interview being interrupted. Perhaps the most important concerns about interviews centre on the effect of the interviewer (Weiss, 1975) - the interviewer may have a direct effect on the interviewee. For example, the social or ethnic characteristic of the interviewer may affect the responses - an interviewer who is found attractive by the respondent may be more likely to obtain answers which the respondent believes will please the interviewer; by clarifying questions the interviewer may distort the meaning of the question to the interviewee; by interpreting the responses the interviewer may distort the meaning intended by the interviewee. Whilst the extent of these effects can be tested by tape-recording of interviews this adds further cost to the data interpretation.

Face-to-face interviews have the specific advantage of allowing the interviewer to clarify issues not only in response to verbal cues, but also in response to non-verbal cues. Conversely, face-to-face interviews are particularly susceptible to the effects of the appearance of the interviewer. Telephone interviews are considerably less costly than individual face-to-face interviews, principally because of the reduction in travel costs. Telephone interviews also reduce the risks of bias caused by the appearance of the interviewer. The major disadvantage of telephone interviews is the absence of any non-verbal cueing to the interviewer. The other disadvantages are the risk of the respondent not being available (or feeling that the timing is intrusive), the risk that the respondent is a substitute for the desired respondent, and the difficulty associated with questions which require the person to choose among various optional responses (as it is not possible for the interviewer to give visual examples of possible responses).

In summary, the major differences between these two types of interview lie in their cost and in the importance of face-to-face contact. Undoubtedly, on a cost basis alone, telephone interviews would be the favoured method. However if it is important to be able to tease issues out in some detail non-verbal cues given by the respondent will take on a greater importance in ensuring both that the question has been understood and that a full answer is being given.

One derivative of the face-to-face interview which helps to contain costs is the group interview (Frey and Fontana, 1991). In their review Frey and Fontana comment that group interviews are more resource-efficient but also that, because the results are "polyphonic" (i.e. arise from contributions from multiple responses), the effect of the interviewer on the interviewees is lessened (Frey and Fontana, 1991). The interviews can be highly flexible, and the group can work together to produce a "group response" in which ideas are moved around and modified (or "corrected" (Schatzman and Strauss, 1973)), thereby reducing the need for the interviewer personally to interpret the responses. Disadvantages include a requirement for the interviewer to be sensitive to group dynamics, and the risks that respondents will feel pressurised to conform, that conflicts may arise and require interpretation or resolution, and that the interviewer can still bias the outcome (particularly if the interviewer is, or becomes, a member of the group).

A number of specific methods for obtaining the views of groups of people have been described (Merton et al. 1956; Van de Ven and Delbecq, 1972; Linstone and Turoff, 1975; Stewart and Shamdasani, 1990). 'Brainstorming' groups aim solely to generate

ideas (Stewart and Shamdasani, 1990). 'Delphi' groups consist of groups of experts in a particular field; they do not meet face-to-face, but are sent postal questionnaires about the area of interest, which are returned to a collating panel who assess the views and then feed those views back to the participants for further response; this process is repeated until consensus (or stalemate) is achieved (Linstone and Turoff, 1975). 'Focus' groups consist of members under the direction of a moderator who maintains the direction of the group within the particular focus (Merton et al. 1956). The 'nominal' derivative of the focus group is designed to generate ideas, and then to discuss those ideas in a highly controlled way, and then to rank them (Van de Ven and Delbecq, 1972); it is particularly designed both to overcome the problem of dominant participants controlling the interview and to allow the group to rank ideas.

These four methods have been contrasted by Gallagher et al. (Gallagher et al. 1993). In brainstorming groups the feasibility of the ideas is not considered. In focus groups ideas are explored in greater detail; this results in a greater risk that certain members will dominate the thinking of the meeting - in this method the skill of the interviewer as group facilitator is of particular importance. Conversely, the nominal group derivative of the focus group uses a highly structured format to rank ideas; it is of less value when the intent of the interview is exploratory, but the level of control does reduce the risk of dominant members controlling the group. In the Delphi method the face-to-face component (and the associated advantages) is lost.

Questionnaires:

Postal questionnaires are the cheapest way of administering questions; the risk of bias being introduced as a result of the appearance or voice of the interviewer is minimised.

Conversely questionnaires are often returned in an unusable format (as a result of absent, illegible or invalid responses); it is also not usually possible to deal with problems that arise for individual respondents in completing the questionnaire. The biggest drawback with postal questionnaires centres on the difficulty of ensuring high response rates to minimise the risk of bias in the results. A number of ways of improving response rates are available: there should be a covering letter which emphasises why the study is important, why that person's response is important, and how the results will be used (Dillman, 1978); stamped envelopes rather than business envelopes should be used (Armstrong and Lusk, 1987); and follow-up questionnaires should be sent to non-responders (Dillman, 1978). The evidence on the effect of the length of the questionnaire on response rates is conflicting (Dillman, 1978; Yu and Cooper, 1983), although it is recommended that the questionnaire length be kept below ten pages (Streiner and Norman, 1995).

Question and answer formats:

Information can be sought by using either 'closed' or 'open' questions (Streiner and Norman, 1995). Closed questions are those which seek one of a number of pre-determined answers - for example 'yes' or 'no', or a number from a range of numbers. Open questions do not seek a pre-determined answer - the respondent can give as much or as little information as possible. The principal advantage of open questions over closed questions is that the respondent is not constrained by predetermined answer formats. The principal problem is that data handling is considerably more complex when the options for answering are not constrained in any way.

For closed questions two main types of response format are available (Streiner and Norman, 1995) - categorical formats (in which the respondent is asked to indicate a response in a "yes or no" form), and continuous formats (in which the respondent is asked to indicate a response on a continuous scale). In choosing between these formats the crucial determinant is whether the response is truly categorical or not - namely whether the answer to the question is simply a "yes" or "no", "true" or "false". For most attitudinal and behavioural issues, responses are rarely truly categorical and continuous answer formats should be used. If a categorical response format is used when the answer is likely to be continuous, three difficulties may arise (Streiner and Norman, 1995): different respondents will have different ideas about what constitutes a positive response; the limited choice of responses will constrain responses (e.g. a positive response will include a range of responses from just positive through to strongly positive); and, as a consequence of this second problem, the instrument becomes less efficient (reducing a continuous variable to two categories results in a considerable increase in the number of responses required to show an effect (Suissa, 1991)).

The most simple version of a continuous scale is the visual analogue scale in which a line of fixed length is drawn between anchors at its two extremes; respondents put a mark at the point on the line that indicates their response. The major alternative is an adjectival scale in which additional descriptions are placed at intervals between the two extremes. Although both formats run the potential risk of introducing response bias (respondents may perceive that a particular response is desired) and of encouraging halo effects (i.e. all items are rated equally on the basis of a global impression without sufficient attention being paid to individual items) and central effects (i.e. respondents rarely pick a response at the extreme), Guilford concludes that, because of the need to minimise the risk of

presenting an "equivocal conception of the continuum" of the response, adjectival scales are probably preferable to visual analogue scales (Guilford, 1954).

Adjectival scales can be further developed by splitting the continuum into multiple specific categories, thereby producing a multi-point categorical scale (Streiner and Norman, 1995). In such a scale the respondents indicate which adjective most nearly corresponds to their response; it is, in effect, a method of bringing a continuous scale into categories. The major advantage of this approach is that the descriptions help respondents to gauge their response along the continuum; the disadvantage is that, however many categories are offered, the categorisation of a continuous response introduces the risks associated with categorical scales (p.86).

When using multi-point adjectival scales, Streiner and Norman offer the following principles in the selection and development of a suitable scale (Streiner and Norman, 1995): firstly, based on the relationship that has been found between the number of categories in an adjectival scale and the reliability of the results, if categories are to be used there should be at least five; secondly that, whilst a unipolar scale may use odd or even numbers of categories, a bipolar scale should use only an odd number; thirdly that descriptors at the ends of scales are probably more important than descriptors for all categories (Wildt and Mazis, 1978); and fourthly that, if numbers are placed alongside descriptors, only positive integers should be used (Schwarz et al. 1991).

Of the multi-point adjectival response formats perhaps the best known is that attributed originally to Likert (Likert, 1952) - namely a multi-point categorical scale based on a

simple continuous axis between extremes of agreement. This is a simple scale which has been used widely.

Conclusions:

A number of formats for seeking views, and for rating answers have been discussed above. My view is that the approach chosen should depend entirely on what information is being sought. Where a large number of views are being sought, a postal questionnaire is likely to prove more efficient provided the problems of response rates can be adequately addressed. Conversely, if it is likely that there will need to be considerable probing to clarify the views of respondents, interviewing techniques are likely to be more helpful. If the views of large numbers are needed, and considerable probing is also likely to be needed, group face-to-face interviews may prove to be most useful.

Similarly, when choosing answer formats the proposed nature of responses will need to be considered and an appropriate format selected; if an adjectival scale is desirable, guidance for the selection or development of such a scale has been offered. In section 4.2 for each of the studies in which views are to be sought the selection of method and answer format will follow the principles outlined in the above section.

4.1.2 Testing properties

It is proposed that a number of properties of a trainer's report should be tested: overall validity, inter-rater reliability and feasibility. These properties are concerned with how the trainer's report form actually works. Consequently, they can only be answered by testing the report form in conditions that are either real, or mimic reality as closely as possible.

Overall validity focuses on whether the instrument measures what it purports to measure - in this context, does the instrument allow the selection out of doctors who are not yet ready for independent practice? There are two components to this issue. The first is whether or not the report form allows the selection out of any doctors; the second is whether or not the doctors selected out are the correct group of doctors (i.e. those who are not yet ready for independent practice). The latter of these is essentially an issue of predictive validity and is subject to the constraints outlined in chapter three, the conclusion being that a test of predictive validity was currently not possible (p.73). The overall validity study will therefore need to focus on an exploration of the degree to which the use of the report form does allow the separation out of a group of doctors - that is, its discriminatory power.

Inter-rater reliability concerns the degree to which assessors agree on their assessment of a single assessee. A field test that includes a test of inter-rater reliability will need to allow assessments to be made by more than one assessor. However, if the field test is to mimic as closely as possible the usual setting in which a trainer's report is to be used (i.e. one in which the trainee would be assessed by a single trainer) it will not be possible for the usual performance of the trainee to be observed by large numbers of assessors. A suitable compromise would be for two trainers to observe the usual performance of the trainee in the usual setting; inter-rater reliability would be based on the degree to which pairs of trainers agreed on their assessments.

Feasibility is concerned with what is possible and practical. It is therefore essential in the testing of feasibility that the setting of the field test should mirror as closely as possible

the setting in which the trainer's report would be used if it were to be adopted widely. In particular, there should be no additional input to the assessors or the assessees in the field test which would not be used in the usual setting - for example, there should be no specific training for trainers unless such training is to be used for all trainers in the long-term. The test of feasibility will need to seek information about the implementation of the trainer's report from those involved in the field test; this should include the seeking of information not only from trainers but also from trainees about the use of the report form. Because this component is a test in which information is sought, the methods for seeking information analysed in the previous section (p.81-8) will be applicable to the feasibility component of a field test.

The final proposed test of the trainer's report is that of content validity. For content validity it was proposed in the conclusion of chapter three that there should be a review of the contents of the report from the perspective of an alternative group of key players from that used in the original selection of content. Because such a proposal requires the seeking of information, again the methods for seeking information analysed earlier are applicable.

Conclusions:

Within the strategy for testing a new trainer's report four properties are to be tested. Of these, two (feasibility and content validity) require views to be sought and will require the selection of one of the methods outlined in the previous section of this chapter. The testing of feasibility will need to form part of a field test in which the testing of the other two properties (discriminatory power and inter-rater reliability) is incorporated. The

principal requirement of the field test is that the conditions should mimic reality closely, although it will be necessary to allow assessment of the trainee by two trainers.

In the following section, the methods selected for the five research studies are described. For each study the rationale for the exact choice of method is explained; this is then followed by a detailed report of the method used.

4.2 Detailed methods

4.2.1 Study 1: Determining an appropriate structure for a trainer's report

General issues

The aim of this study is to determine an appropriate structure for the trainer's report. In chapter three it has been argued that this study should, in particular, use the experience of trainers to draw conclusions about how a trainer's report would need to function and, through that process, to draw conclusions on the most effective structure for a new trainer's report, thereby promoting feasible assessment (the premise being that, by incorporating the views of trainers, feasibility is likely to be enhanced). The results of this study, in conjunction with theory and experience from elsewhere should enable the development of a report form that promotes predictive validity (in particular by encouraging assessment that is based on the observation of performance). It is also intended that this process of consultation with trainers should be used to ensure that the rest of the research programme is likely to answer the most important issues - that is, it will enable the validity of the research proposals to be tested.

To achieve this aim the chosen methodology needs to ensure that there should be a discussion of as many options as possible; this requires that the views of a broad range of

trainers should be sought (to maximise the range of options being considered). Because it is predominantly an exploratory exercise the ability to be able to tease out issues in some depth becomes very important. This need for discussion rules out the use of a postal questionnaire survey as an appropriate methodology for this study, leaving a choice between telephone or face-to-face interviews.

In the analysis of methods of seeking views contained in the first section of this chapter it was concluded that the combination of the need to include the view of a broad range of interviewees and to enable qualitative interpretation of the responses is most likely to be met by means of interviews using the focus group technique (p.83-4). In the particular context of general practice natural groupings of trainers already exist. Trainers' groups provide a format for mutual support and for discussion of educational initiatives. These groups could be used as the focus group for the interview.

Because the prime aim of this study is to explore the views of trainers on the structure of a trainer's report, open questions should form a substantial part of the interview; the questions should seek their views on the structure of the report form, allowing exploration of potential options in doing so.

Study methods

Interview groups

During mid-1994 trainer's groups in five National Health Service (NHS) Regions were asked, by letter, if they would be prepared to allow one of their regular trainers' group meetings to be used for the purposes of an interview about the design of a trainer's report for summative assessment. The regions were selected to ensure a wide

geographical spread and to include groups from both rural and urban areas. To do this the Directors of Post-graduate General Practice Education responsible for the Oxford, West Midlands, Northern Ireland, South-East Scotland and North-West England regions were approached. From the lists that they provided, 17 trainers' groups were approached. These represented all nine from the Oxford region and two, selected at random, from each of the other four regions. This combination of groups was chosen to maximise the efficiency of the study (by increasing the number of local groups) whilst still ensuring a broad geographical spread of groups. Groups failing to respond to the letter received a follow-up letter one month later.

Structure of interview

At the beginning of each meeting a brief introduction to the proposed summative assessment process was given, along with a description of where a trainer's report would fall within this process. The group was assured that any comments that they made would not be attributable to the group.

Questions

Questions were chosen to examine three issues. To provide an indication of the way in which trainers currently operate, and to allow them to analyse perceived strengths and weaknesses of such an approach, trainers were asked "what, if anything, do you currently use as the basis for a report by the trainer?". To obtain views on the design of the form trainers were asked "what do you see as the pros and cons of your current approach?" and "what do you feel would be the most appropriate design for a structured trainer's report for summative assessment?". It was intended that this would enable them to become innovative without losing any sense of what might be realistically achievable (i.e.

to promote feasibility in any suggestions). The open form of question was designed to minimise the chances that trainers would feel that there was a specific "desired" answer (i.e. to encourage discussion of as many options as possible). To establish whether trainers might feel able to express concerns about a trainee, and to assess how a structured report form might facilitate that process (i.e. to see if a structured report form might offer any significant advantages over the system currently in operation) trainers were asked "as trainers, have you ever had concerns about signing up a trainee on the current VTR1 form? If so, how might a structured trainer's report have helped in making this decision?".

During the interview notes were kept and transcribed into minutes within 24 hours of the interview. At the end of the interview a chance was given for the group to make any comments or ask any questions about summative assessment in general, but this component was not used as part of the minutes recorded.

4.2.2 Study 2: Selecting appropriate contents

General issues

The aim of this study is to select appropriate content for a trainer's report as part of a summative assessment blueprint. It has been argued in chapter three that this study should follow the concept of a service-based assessment blueprint. In this study the attributes required by practitioners to deliver the service are to be selected on the basis of the views of one group of key players - general practitioner trainers - through their views on the importance for independent general practice of these attributes. To enable a comprehensive assessment blueprint to be developed opinion would be sought on matching appropriate instruments to those attributes. In addition, whilst some

information on the likely frequency of trainees failing a pilot summative assessment process has been published (Campbell et al. 1993), a consultation exercise offers the opportunity to check the validity of this estimate; a secondary aim of this study is therefore to obtain an estimate of the frequency with which trainers have concerns about the performance of trainees.

The need to maximise content validity is most likely to be met if the views of a large, representative sample of trainers from across the U.K. is included in the selection. This has the benefits both of reducing the risk of the report being considered to have a local flavour (a problem with the North Thames (West) report form (Rhodes and Styles, 1995)) and of increasing any benefits that might come from trainers feeling that they have had some involvement in its development (Rhodes, 1998). This requirement is most efficiently addressed by using a postal questionnaire survey. If so, the methods for maximising response rates outlined in the first section of this chapter (p.85) must be borne in mind.

In selecting the answer formats for the questions, for the component of this study in which the contents are to be matched to assessment methods the choice is relatively straightforward. Four methods have been suggested (Joint Committee on Postgraduate Training for General Practice Working Party on Assessment, 1992a); the logical solution is to use closed questions using a categorical answer format based on choosing one of the four proposed methods. Additional information can be obtained by providing a response format which allows the respondent to indicate any other potential methods that are felt to be important (i.e. providing an open response format to an otherwise closed question).

For the component in which contents will be selected on the basis of the importance for independent general practice ascribed by trainers there are two main options. The first is to use open questions in which trainers are asked what they believe should be contained within the report. This approach allows trainers the freedom to include whatever they believe to be important but it does require sophisticated data-interpretation techniques, in particular to ensure that the interpretation of the response given is actually that intended by the respondent. The second option is to use closed questions in which respondents are asked to rate the importance of elements of general practice already judged to be important (Statement by a working party of the second European conference on the teaching of general practice, 1977; Oxford region course organisers and regional advisers group, 1985; Centre for Primary Care Research, University of Manchester, 1988). This second approach does simplify the data interpretation but it may stifle new ideas. On balance, because of the desire to involve large numbers of trainers it is important that the interpretation of large amounts of data is straightforward; this would be most easily achieved by using closed questions based on predetermined elements of general practice. To prevent the exclusion of previously undescribed ideas, open questions inviting additional suggestions could also be included.

Answer formats which allow respondents to indicate their view of the importance of each of the elements included will need to be selected. In this instance respondents will need to be able to indicate a judgement which falls somewhere along a scale between the extremes of "not important at all" and "crucial" for independent general practice. In this particular study an additional issue arises - that of the positive skew bias (Streiner and Norman, 1995). Where respondents are being asked to rate importance for elements that

have already been described elsewhere as being important for general practice, it is likely that if a simple visual analogue scale were to be used the majority of responses would be bunched near to the "important" end of the scale. One way to provide more detail on these responses is to provide an adjectival scale which focuses on the "important" end of the continuum. To consider this possibility, it is suggested that this study should include a pilot exercise in which two answer formats (one a simple scale, and one a skewed scale) are considered.

Before detailing this study it is important to clarify the medico-political context in which it was undertaken. This study began in April 1994; all data had been collected by December 1994. Following the publication of the working paper from the JCPTGP (Joint Committee on Postgraduate Training for General Practice, 1994), it was becoming common knowledge that summative assessment was being planned, but the details were still fairly sketchy. Trainers were becoming aware that it was likely that they would be required to submit a formal report on their trainee, but the specific nature of the assessment tools to be used (other than the multiple choice questionnaire) was still under debate. In particular, there was considerable debate around the format of the observation of practice, with discussion centring on the use of either simulated surgeries (Rethans et al. 1991) or video-taped consultations (Campbell et al. 1995a). This has a particular bearing on the trainer's report because, whilst simulated surgeries would allow the assessment of many clinical skills, the video-taping of consultations provides very limited scope for the assessment of such skills. The final decision about which technique to adopt was not taken until early 1995. Consequently trainers involved in this study could not be sure that the assessment of clinical skills would be reliably undertaken by a component of summative assessment other than the trainer's report.

Study methods

The study was based on a postal questionnaire survey of a random national sample of general practitioner trainers.

Questionnaire content

The content of the questionnaire was based on three previously published documents - the two statements of the range of qualities required of independent general practitioners (Statement by a working party of the second European conference on the teaching of general practice, 1977; Oxford region course organisers and regional advisers group, 1985), in conjunction with the one established assessment format (Centre for Primary Care Research, University of Manchester, 1988). The Leeuwenhorst statement (Statement by a working party of the second European conference on the teaching of general practice, 1977) describes the educational aims for the work of the general practitioner in terms of 23 broad statements. The Oxford Region Priority Objectives (Oxford region course organisers and regional advisers group, 1985) describe 5 broad educational areas (care, communication, organisation, professional values, and personal and professional growth) which are subdivided into 45 specific educational objectives for the general practice training year. The Manchester Rating Scales (Centre for Primary Care Research, University of Manchester, 1988) consist of 23 main scales, each of which is divided into subscales; in total there are 165 subscales. In developing the questionnaire all main scales of the Manchester Rating Scales were included; all main sections of the Leeuwenhorst statement and the Priority Objectives were also included. Because of the strong possibility at the time of the development of this questionnaire that the observation of practice would take the form of video-taped consultations (which could not ensure that all clinical skills would be assessed) all seventeen subscales of the

Manchester Rating Scales which describe specific clinical skills were also incorporated into the questionnaire.

In total, questions on 75 elements of general practice were included. For clarity the elements were grouped into the five categories used in the Oxford Region Priority Objectives (Oxford region course organisers and regional advisers group. 1985) - namely "patient care" (42 elements), "communication" (5 elements), "organisation" (14 elements), "professional values" (8 elements) and "personal and professional growth" (6 elements).

Questionnaire design and piloting

For each element of general practice two questions were asked. The first asked the trainers to rate the importance of that element for independent general practice. The second question for each element of practice asked trainers to indicate which of five possible methods of assessment might be used to assess this element; the options given were: "written exam", "external observation", "trainee project" and "trainer's report"; a category entitled "other" was also included. Trainers were asked to tick all the boxes that would, in their opinion, offer suitable assessment methods for that attribute; if they ticked the "other" box they were also asked to describe what technique they would advocate.

Towards the end of the questionnaire questions seeking basic demographic data on each trainer were included. To fulfill the subsidiary aim of this study, each trainer was asked whether or not he or she had ever considered not signing the VTR 1 form on a trainee.

Finally respondents were asked to list any additional questions that they wished to see included in the trainer's report.

The first draft of the questionnaire was piloted to 6 general practitioners who had a strong commitment to training within the Oxford region. They were asked to make general comments about the questionnaire; they were particularly asked to comment on which of two possible answer formats allowed them to indicate most accurately their view about the importance of a particular item. The first format was a traditional Likert scale (Likert, 1952) using a five-point response scale from "strongly disagree" to "strongly agree". The second consisted of a five-point adjectival scale which focused on the "important" end of the continuum (i.e. 1 = "fairly important" up to 5 = "crucial"). All respondents to the pilot questionnaire favoured the second format. This resulted in a second draft which was then forwarded to one randomly selected trainer from each of the 20 regions that had provided lists of trainers at this stage (July, 1994). They were asked specifically to comment on the questions used and on the accompanying letter that was to be sent with the questionnaire. Thirteen replied and their replies were used to form the final draft. An example of the final question format is given below, and the full questionnaire is provided as appendix 4.1.

Figure 4.1: Content study - example of question format

1. The doctor can recognise common physical, psychological and social problems

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

A covering letter explained the reason for undertaking this study, the basis of the questionnaire, and exactly what each of the assessment methods might entail (including an explanation of the options of either video-taped consultations or simulated surgeries for the observation of performance).

Sample size and structure

All Directors of Post-graduate General Practice Education in the United Kingdom were approached in June 1994 to ask if they would be willing to provide lists of the trainers in their regions. A total of 27 advisers were approached (24 civilian, 3 armed forces) of whom 26 agreed to be involved. This resulted in the names of 3335 trainers being available for inclusion in the study out of a total of 3447 trainers in the United Kingdom at the time of the study (information obtained from all 27 regions).

The sample size calculation was based on the following (Mant and Yudkin, 1993):

$$n = \frac{1.96^2 Np(1-p)}{(1.96^2 p(1-p)) + \Delta^2 N}$$

where n is the sample size, N the total number from which the sample is taken, p the estimate of the proportion of responses falling within the desired category (the maximum sample sizes being based on p = 0.5) and Δ is the desired range. Using N = 3447, p = 0.5, and Δ = 0.025, n = 1063; this is the maximum number of responses needed to ensure a 95% probability of the result lying within 2.5% of the true result. If p = 0.7, n = 939. Based on a response rate to the questionnaire of 75% the sample sizes required would be 1417 and 1252 respectively. A "worst case scenario" was then considered. If a sample size of 1300 was chosen, and only 60% responded, and the true value of p = 0.5, the

value of Δ would still be 0.031 i.e. the result would have a 95% probability of lying within 3.1% of the true result. Conversely, in the best case (i.e. 80% response rate, p = 0.8) Δ becomes 0.020. A sample size of approximately 1300 was therefore chosen. This represents 38.98% of 3335; for simplicity a sample of as near as possible to 39% of 3335 was chosen. A random number table (Bland, 1987) containing 500 numbers was used to select 195 of every 500 trainers; a total of 1298 trainers were selected of whom 2 were duplicates, leaving a final sample of 1296. Whenever the random selection resulted in the selection of one of the trainers who had been approached in the pilot phase, the next name on the list was selected.

Questionnaires were sent initially in September 1994 with follow-up questionnaires to non-responders after 4 weeks and again after a further 5 weeks.

Data analysis

Responses were entered into the EPI-INFO software package. For each of the 75 elements under scrutiny two calculations were made. Firstly the number and percentage of responses in each of the five "levels" of importance were calculated; because all adjectival scales risk "centre bias" (namely that respondents will tend not to rate at the extremes of a rating scale) (Streiner and Norman, 1995), when considering the relative importance of elements the numbers and percentages for categories 1 with 2 and 4 with 5 were combined. Secondly the number and percentage of responses for each assessment method were calculated. Because trainers could indicate more than one possible response to this question, the following responses were considered specifically for each item to enable comparison of different ways of analysing the data: the proportion

indicating the trainer's report alone; the modal response; and the proportion indicating a response that did not include the trainer's report at all.

The denominator for the calculation of all percentages was the total number of responses for that question; for the importance questions the lowest number of responses was 904 (92.8% of 974), whilst for the assessment methods questions the lowest number of responses was 850 (87.3% of 974) (this being the response rate for the assessment question for the item judged least important by trainers).

Confidence intervals for the percentages were calculated using the Confidence Interval Analysis package (Gardner et al. 1992). Comparisons between respondents and non-respondents were made using the chi-square test for proportions and the standard error of difference between means (Bland, 1987).

4.2.3 Study 3: Assessing content validity

General issues

The aim of this study is to assess the views of another group of key players on the validity of the content of the proposed trainer's report. This proposal is included to minimise the risk of introducing bias to the report through the exclusive use of one group of key players in the selection of content; the sample used in the content selection study (i.e. general practitioner trainers) provides a limited perspective of the attributes required for independent general practice. An alternative view of the validity of the contents could provide evidence about whether or not such views are shared by others with a stake in assessment.

Content validity is concerned with the "adequacy of sampling of the specified universe of content" (Fabb and Marshall, 1983). This phrase can be interpreted in two ways: firstly, the term "sampling" can be considered to focus on the issue of the selection of contents for the assessment process; secondly, the term "sampling" can be considered to refer to the process of assessment i.e. it concerns the adequacy of the sampling of behaviours within the assessment process itself. Of these two interpretations it is the former that constitutes content validity; the latter is one aspect of the overall validity of the assessment.

In selecting an appropriate method for this study two options are possible - either to obtain an alternative *de novo* view of the contents, or to seek the views of an alternative group on the contents already selected by trainers. Whilst both approaches would provide information about the views of the alternative group, each approach has particular problems associated with it. The principal problem with the former approach is that a different combination of attributes may well be selected as important; if this were the case, how is a decision to be made as to which view should prevail? With the latter approach the principal problem is that the study group is effectively only being asked to ratify a decision that has already been made. On balance, in my view, for a study that is attempting to assess the validity of the proposed contents of the report the latter approach is acceptable and much less likely to yield findings which are very difficult to integrate with the findings of study two without accusations of the introduction of bias. Consequently I would argue that this study should be undertaken by a direct comparison of the views of another group of key players on the degree to which they agree or disagree that the proposed contents are elements that are needed for independent general practice and can be assessed by means of a trainer's report. This could be effectively

achieved using a questionnaire survey, based on closed questions with two Likert response scales (one for importance and one for assessibility) using questions based on the attributes selected in study two.

One particular risk of questionnaires with multiple items using the same response scale is that of acquiescence bias - namely the tendency for respondents to give positive responses to all questions (Couch and Keniston, 1960). One way to try to avoid this is to have "equal numbers of items keyed in the positive and negative directions" (Streiner and Norman, 1995). The risk of such an approach is that the respondent may become confused and, unless attempts are made to measure the strength of this effect, it will be impossible to know how great this effect is. An alternative approach is to leave the response formats to all questions keyed in the same direction, but to include some items which would be expected to result in a substantially different response. Within the questionnaire proposed for this study, this could be most easily achieved by including not only all the proposed items for the trainer's report (each of which might be expected to associated with high levels of agreement to their inclusion) but also some items which might be expected to be associated with low levels of agreement. The items judged least important by the trainers in study two provide suitable items for such a test of acquiescence bias.

The views of those particularly affected by the assessment process are likely to prove helpful if the credibility of an assessment process is to be maintained. In study two the view from one end of the assessment microscope, that of the assessors (i.e. the trainers), was considered. A useful alternative perspective is likely to be offered from the other end of the microscope - namely the assessed (i.e. the trainees). Seeking the views of

those in training might be subject to the criticism that they might be considered not to have had sufficient experience to be able to judge what was most important for independent general practice. An alternative is to consider the views of those who have recently completed vocational training. Because these doctors have completed their training they might be expected to have a broader view about what was important for independent general practice than doctors still undertaking training. In addition, although their views might be, at least to some extent, influenced by their trainers, it might be expected that they had had a greater chance to form independent views about importance and assessibility than doctors still under regular supervision by a trainer. For these reasons this group would seem to offer a suitable sample for a validation study.

Study methods

This study examines the validity of the proposed contents of the trainer's report by measuring the extent to which doctors who had recently completed vocational training agree that the proposed contents reflect activities needed in general practice and that it is reasonable to assess the proposed contents by means of a trainer's report.

Questionnaire

A questionnaire was drawn up using the items contained in the draft trainer's report (appendix 4.2). In this study, one of the items selected in study two is split into two items; this occurs because, during study four (which was conducted just before this study) there was agreement that understanding of the meaning of one item ("the doctor undertakes appropriate examination with appropriate consideration of the patients needs and feelings") was considerably improved by splitting it into two component parts ("the doctor undertakes appropriate examination (including investigations)" and "the doctor

undertakes examination with appropriate consideration of the patient's needs and feelings"). In addition, the two items judged to be least important by trainers were also included (namely "the doctor has an understanding of the basic methods of research as applied to general practice" and "the doctor is able to use the laryngoscope proficiently and to interpret the findings made").

For each item respondents were asked to indicate their response to two statements by means of five-point Likert scale (Likert, 1952). The two statements were: "this is a skill that is needed in general practice" and "it is reasonable that this skill is assessed by means of a trainer's report". An example of the question format is given below and the full questionnaire is shown as appendix 4.2.

Figure 4.2: Content validity study - example of question format

2: The doctor is able to examine each system and each organ proficiently

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

At the end of the questionnaire the age and gender of respondents was sought. Respondents were also asked for any comments that they would like to make about the trainer's report with space being left for freetext comments.

A covering letter was sent with each questionnaire explaining the reason for the study. The letter explained what was required of respondents - namely, for the first question,

based on their experience in general practice their view on whether the item reflected a piece of knowledge, a skill or an attitude that is needed in general practice; and for the second question whether it is reasonable for this item to be assessed by means of a report provided by the trainer, based on assessment undertaken during the training year.

The first mailing was sent in May 1995. Two subsequent mailings were sent at five week intervals to non-responders. Addresses of non-responders to the first mailing were checked against the Medical Register (General Medical Council, 1994).

Sample

The sample size was calculated based on an *ad hoc* estimate of the proportion agreeing with the outcome of study two of 0.9 (based on the assumption that the views of those who had recently completed vocational training were likely to be broadly similar to those of trainers) with an acceptable level of tolerance for the results of the study of 0.05. This required 138 responses (Mant and Yudkin, 1993). Based on a response rate of 65% this would require the sending of 212 questionnaires. Of all 1933 doctors who had received certificates of completion of vocational training from the JCPTGP during 1994 (Joint Committee on Postgraduate Training for General Practice, 1995) 220 (11.4%) were randomly selected for inclusion in the study.

4.2.4 Study 4: Setting standards

General issues

The aim of this study is to develop minimum standards for each of the attributes selected for inclusion in the trainer's report. On p.29 it was argued that, with suitable safeguards, there remains a place for self-regulation by the professions; consequently this study is

designed to enable members of the profession to set standards. In addition, to promote the feasibility of the trainer's report, this study also seeks views on the ways in which the testing of performance against these standards could be realistically undertaken - in particular how performance can be best tested, and who is in the best position to observe it. It has already been argued in chapter three that predictive validity is most likely to be achieved if direct observation of performance can be encouraged (p.61). If this requirement is to be met all elements contained in the trainer's report should be amenable to direct observation of the trainee by the trainer; if the view of the standard-setters is that all or most of the attributes can be tested by direct observation of performance it is likely that the predictive validity of the report form will be considerably strengthened.

It is important that the method of standard setting is fair (Dauphinee, 1994) and unbiased (Bowmer, 1994), including the need for the "standards to be set by an adequate number of judges who are knowledgeable, some of whom are experts or leaders in the field" (Bowmer, 1994). It is also important that appropriate standards are set - absolute standards (i.e. criterion-referenced standards) rather than relative standards (i.e. peer-referenced standards) (p.49). It has already been argued in chapter three that the standards needed for this report are absolute standards set at the minimum level for independent general practice (p.54) - standards described by Kane as those that "provide a clear and credible basis for differentiating good performance from bad performance" (Kane, 1992). The probability of "unclassification" must be minimised (Crocker and Algina, 1986) - that is, it should be possible to place each assessee clearly in one category or another; in the case of minimum standards this is most effectively addressed by using a single classification at the level of minimum standard (i.e. a pass or fail decision).

Methods of setting absolute standards for written examinations of competence are well established (Ebel, 1979; Angoff, 1971). These involve groups of experienced teachers considering the proportion of candidates, at this particular point in their training, who might be expected to provide the correct answer. Whilst there is currently very limited experience of setting absolute standards for a method such as a trainer's report (Rakowski, 1990; Preece, 1993; Rhodes and Styles, 1995) a similar system, in which a group of experienced judges set consensus standards, could be used. Although there is a risk that the standards can be set at an unreasonably high level (Norman et al. 1985; Rethans et al. 1991), if it is possible to set minimum standards for written tests in this way then it should be equally possible to develop suitable minimum standards for a trainer's report in a similar way.

Bowmer (Bowmer, 1994) recommends that at least ten people should be involved in the exercise, although "it may be possible to use smaller groups if several different groups are working on different parts of the test"; Crocker and Algina argue that "multiple samples of judges" are needed to minimise bias (Crocker and Algina, 1986). Bowmer also recommends that those involved should be knowledgeable in the field of the examination and some should be experts (Bowmer, 1994). He makes two further recommendations - firstly that those responsible for the development of the assessment tool should not be involved in the setting of standards, and secondly that, when entry into a profession is the focus, the group should include educators who are familiar with the training process and the level of ability of those completing training (i.e. for this report, general practitioner trainers).

If minimum standards are to be developed using a method in which groups develop consensus standards a process that supports the development of consensus is needed. In the development of standards for clinical performance review, two main approaches have been described. The first is the consensus conference (Glaser, 1980). The second is the Delphi technique (Linstone and Turoff, 1975). In the consensus conference delegates work face-to-face in groups analogous to focus groups supported by facilitators. It has the advantage of allowing fairly rapid development of consensus as areas of agreement can usually be found quickly, and areas of disagreement can be explored at the time of the conference. The process can suffer from the disadvantages of focus groups (p.83-4) - it can be dominated by strong personalities, and the effect of the facilitator is unpredictable. The Delphi technique involves groups who receive postal questionnaires about the area of interest which are then returned to a collating panel who assess the views and then feed them back to the participants for further response. This reduces the risk of the process being dominated by strong individuals, but it does increase the time taken to complete the process and is at considerable risk of a facilitator-effect.

Because experience in this field is limited, whatever method is chosen for this study would be essentially exploratory. Based on the experience of standard-setting for written tests (Ebel, 1979; Angoff, 1971), both of which use the consensus conference, it would seem reasonable to base the standard-setting exercise for the trainer's report initially on a consensus conference. Based on the suggestion that bias is reduced by using multiple samples of judges (Crocker and Algina, 1986) a second phase, using a Delphi method, could be included.

Study methods

The items which form the basis of this exercise were those resulting from the survey described in study two.

The standard-setting was undertaken in two stages. Firstly a consensus conference was held. Secondly, a Delphi exercise was undertaken in which the outcome of the conference was fed back both to the attenders of the conference and also to a number of other experts whose views were felt to be crucial in the setting of minimum standards for independent general practice.

Consensus conference

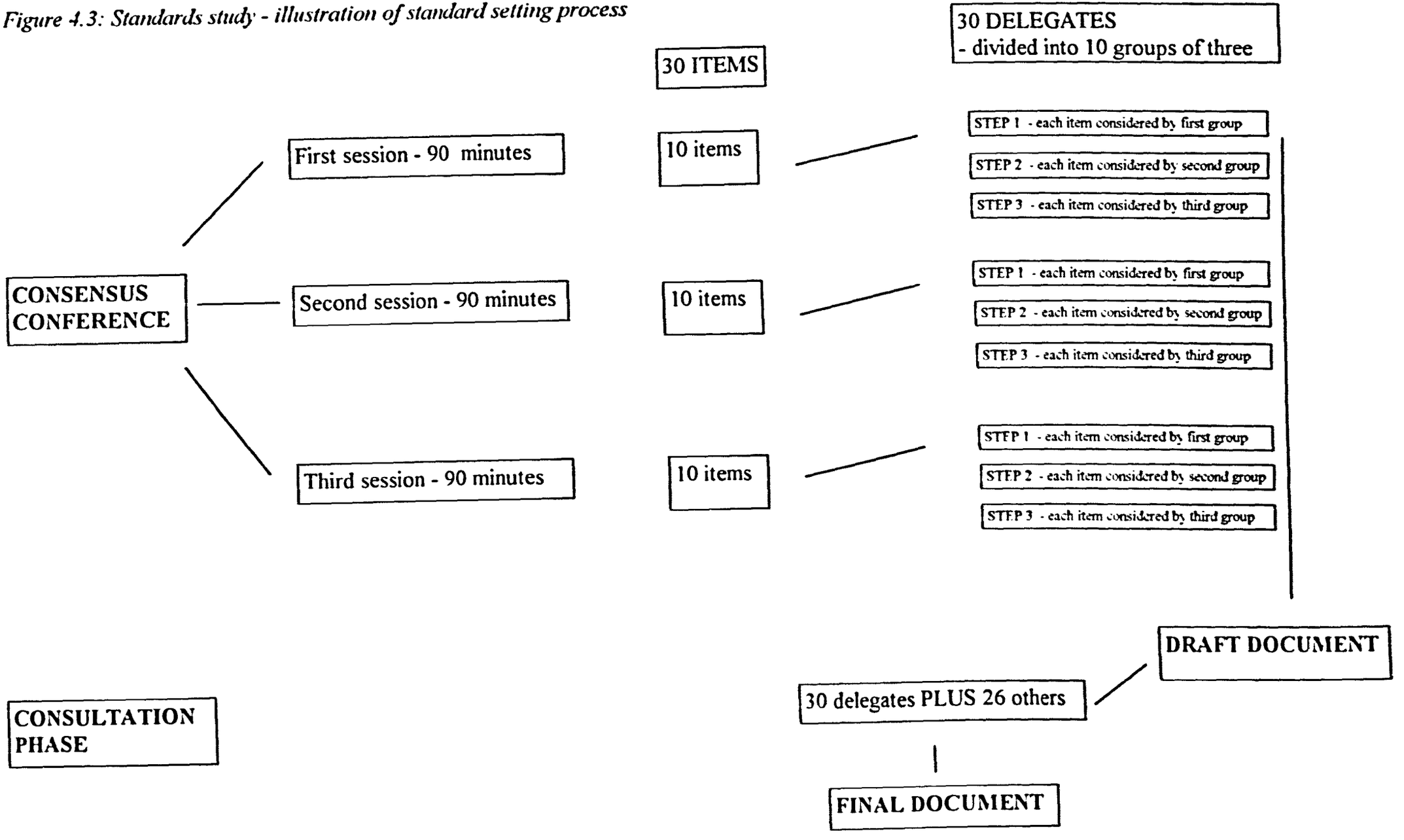
Directors of Postgraduate General Practice Education in the United Kingdom were asked in December 1994 to nominate two experienced trainers from their region who might be willing to attend a consensus conference. The first named trainer was invited by post to attend the meeting. If there was any doubt about whether the first named trainer would be able to attend, the second named trainer was invited. This resulted in 27 trainers attending who represented 21 regions from the United Kingdom (19 civilian, 2 armed forces). A representative of the JCPTGP also attended. Following a consultation meeting with trainee representatives in February 1995 they were also invited to attend. Five expressed interest and two attended.

The conference took place in March 1995. The process of the conference is illustrated in figure 4.3 (p.114). Following a brief introduction the thirty attenders were divided into ten groups of three; no group contained two representatives from the same region. The use of multiple small groups to undertake the standard-setting was chosen to encourage

all members to contribute, to reduce the chances that the standards set might be heavily influenced by one particularly dominant member of one group, and to minimise the chances of delegates getting bored and losing interest by spending too much time on any one standard.

Each of the thirty items was considered by three groups, with each group considering the item for 30 minutes. There was no consistent pattern in the order in which groups followed each other. Prior to the meeting delegates were informed of the aims of the conference and given guidelines on the work involved (appendix 4.3), a list of the nine items they would be covering during the day (along with a list of all thirty items being considered) and an example of the sort of outcome that was being sought (appendix 4.4). The conference itself lasted one day.

Figure 4.3: Standards study - illustration of standard setting process



To fulfill the three aims of the study the groups were asked to develop a group view for three areas. Firstly, they were asked to develop a minimum standard; to encourage a focus on minimum standards delegates were asked to consider "what would constitute a failure" for this item (that is, to focus on the performance of a typical sample of those at the borderline). Secondly they were asked to decide what methods for gathering evidence might provide suitable evidence for the assessment. Thirdly, they were asked to advise as to who, other than the trainer, might be able to provide that evidence.

In order to develop consensus the groups were asked to write their conclusions on a worksheet (appendix 4.5) which was passed on to the next group considering the item until all three groups had written their conclusions on it. When a group was the second group considering an item, it was suggested that they look at the item afresh for most of the thirty minutes, and use the last few minutes to consider the conclusions of the first group and decide if they wished to alter their conclusions as a result. When a group was the third group considering an item, it was suggested that they look at the item afresh for the first fifteen minutes, and then use the second fifteen minutes to consider the conclusions of the first two groups before writing down their own. To encourage this process the four members of the steering committee for the overall project moved from group to group, although they did not take part in the standard-setting itself.

Once the conference was completed the conclusions of the groups were drawn together to produce a first draft of the standards that could then be circulated for consultation.

Consultation phase

Two groups were involved in the consultation phase. The first were all thirty attenders at the consensus conference. They were asked to consider in particular the nine items

that they had looked at in detail on the day, to consider also the other 21 items, and to make any other comments on the draft standards that they wished.

The second group were experts whose views were felt to be essential in the setting of minimum standards for independent general practice; there were 26 in this group. This group consisted of all Directors of Postgraduate General Practice Education (except one who was on the steering committee for this project, and one who attended the consensus conference), the Chairman of Council of the Royal College of General Practitioners and the Chairman of the JCPTGP. They were asked to comment on all draft standards and to make any other comments that they wished.

4.2.5 Study 5: Assessing overall validity, inter-rater reliability and feasibility

General issues

The aim of this study is to assess the overall validity (by means of discriminatory power), inter-rater reliability and feasibility of the report form. All of these properties are to be tested in conditions that mimic as closely as possible the setting in which the report form is designed to be used.

It has been argued above (p.89) that the test of overall validity involves two components - discriminatory power and predictive validity - of which only discriminatory power could realistically be tested. If a trainer's report fails to separate out any doctors then, assuming that there are doctors within the group assessed whose performance does fail to reach the minimum standard, it is failing to measure what it purports to measure. The evidence available suggests that, at the end of training, about 5% of trainees will cause their trainer concern (Campbell et al. 1993). The smallness of this proportion presents considerable problems in the design of a field test - the smaller the predicted frequency of

the property under consideration, the larger the sample size needed to measure the frequency with any accuracy (Mant and Yudkin, 1993). One option to reduce this difficulty is to attempt to maximise the likely frequency of doctors in the field test who may have performance that is close to the pass/fail borderline. This could be done either by deliberately including doctors whose performance is already in question, or by deliberately encouraging assessments to be done early in the training of some trainees i.e. at a point at which their performance is more likely to be below the minimum standard set for independent general practice. Of these two options the former is unattractive as it would require selection of trainees for the field test. This would leave the results of the field test open to the criticism that they could not be generalised to all trainees and their trainers. The most viable option is to include trainees who are still at an early stage in their training. If this is the option chosen, though, it means that the field test can not be undertaken over a full year (because this would return us to the position of only about 5% of the trainees under-performing). In summary, if discriminatory power is to be assessed, the most realistic option is to undertake a field test that involves assessment over a period shorter than the full training year, and that deliberately aims to include some trainees who, by the end of the field test, would not be expected to have completed their training. This approach also has the advantage of allowing the inclusion in the field test of those trainees who are not undertaking single periods of twelve months in a training practice.

Discriminatory power concerns the likelihood that a trainee will fail a particular item on the trainer's report. I believe that it can be divided into two types - the "absolute" value of discriminatory power of the report (or an individual item within the report) concerns the likelihood that trainees completing training will fail (i.e. it is an absolute value for trainees at the end of training); an "indicative" value of discriminatory power provides an

indication of whether or not an individual item, or the whole report, will enable the selection out of any trainees as failing (i.e. it is a measure of whether the item or report will discriminate at all). Absolute values are useful in defining the attributes with which trainees have greatest difficulty at the completion of training (i.e. they provide information about the performance of the trainees) but absolute values can only be derived by considering the results of report forms at the completion of training. Conversely indicative values, which can be derived through the assessment of doctors at any point in their training, provide information about whether the test format does actually work (i.e. they provide information about the performance of the assessment instrument). The aim of this field test is principally to examine the properties of the instrument. Consequently I believe that it is acceptable to consider indicative values determined through a field test undertaken with trainees at various points in their general practice training. This is most simply presented as the proportion of trainees whose performance is assessed as being below the indicated minimum standard.

Inter-rater reliability focuses on the degree to which separate assessors reach the same conclusions. The testing of inter-rater reliability introduces another set of requirements for the field test. A field test that includes a test of inter-rater reliability will need to allow assessments to be made by two assessors. Bearing in mind the nature of the trainer's report, these two assessors should both be trainers. There are three risks associated with this approach: firstly, the approach considerably limits the number of practices that might be approached to be involved; secondly, because only a limited group of practices can be approached, there is a risk that the sample used will not be representative of trainers as a whole; finally, because of the emphasis in the trainer's report on the need for direct observation of the trainee, there is a considerable risk of a field test dominating the training. This last drawback could be reduced by undertaking

the field test over a period that is considerably less than a full training year, but there is no obvious way of overcoming the first two problems in a study designed to measure inter-rater reliability. It will be important in the interpretation of the results to bear in mind the limitations posed by the method used.

There are three ways in which the results of the reliability study could be presented. The first is simply to present the proportion of instances in which the two assessors agree. Whilst these are very simple data to derive there is one particular disadvantage to this approach - namely that, when one result is considerably more likely to occur than the other (in this case a pass would be expected to be much more likely to occur than a failure), a high proportion of agreement might be expected occur purely as a result of chance. A second way is to attempt to remove the effect of chance agreement by using a co-efficient of agreement based on the difference between the observed level of agreement and the level of agreement that might be expected purely as a result of chance alone. This is the basis of the most widely used co-efficient of agreement for situations in which there are only two levels (in this instance pass or fail) - the *kappa* coefficient originally described by Cohen (Cohen, 1960). The third way of presenting data on levels of agreement is to use a "generalisability" coefficient. This coefficient attempts to "recognise and estimate the magnitude of the multiple sources of measurement error" (Shavelson et al. 1989). Whilst, because of its attempt to include all sources of error, this approach initially looks attractive there are two areas that require caution. The first is that, in moving from reliability to generalisability, the gain of having a single coefficient to incorporate all sources of error is balanced by the loss of specific information on the relative contributions of each of these sources of error. Secondly, because of the wish to include all sources of error, multiple different forms of data must be included in the calculation - that is, it is a manoeuvre of considerable complexity. For

this study, for two reasons, a middle course of action, based on the use of a coefficient of agreement, may be the most suitable approach. The first reason is that, for the reasons given earlier (p.63), the most important aspect of reliability is that of inter-rater reliability, a property that can be completely described in terms of a coefficient of agreement rather than being incorporated into a more global generalisability coefficient. The second reason is that, in a study that is already considerably complicated by the desire to assess three properties, the requirement for a yet more complex study is probably not justified by the additional understanding that would be offered by presenting a coefficient of generalisability rather than a coefficient of agreement. It is therefore suggested that this study should be confined to the collection of data that will allow the calculation of a coefficient of agreement. Even with this relatively simple coefficient, problems remain. Because of the way in which it is calculated if there are no entries in any of the four cells (i.e. assessor one, pass or fail, *vs.* assessor two, pass or fail) it is not possible to calculate the *kappa* coefficient; even when there might be 100% agreement, no *kappa* value can be calculated. It is therefore proposed that for all individual items, and for the report itself, a simple indication of the proportion of instances in which the two trainers agreed on the results of their assessments should still be presented.

In the testing of feasibility, the principal requirement is that the setting of the field test should mirror as closely as possible the setting in which the trainer's report would be used if it were to be adopted widely. The most simple way of obtaining a quantitative estimate of the feasibility of completion of items in the report is to calculate the proportion of instances in which the trainers find it impossible to complete the assessment of that item. If the field test is to be completed in a time period that is less than would normally be used, whilst this estimate can not be considered to provide an

absolute estimate of the feasibility of completion (in the same way that absolute values of discriminatory power can not be provided unless all the trainees involved are completing their training at the time of the field test), it would provide an indication of the levels of feasibility. Indeed, because those items that are easier to assess are probably more likely to be completed if there is time pressure on the trainers, it might be argued that the results of a feasibility study undertaken in a period of less than the full training year will provide a particularly strong indication of which items are most likely to cause difficulty for the trainer in their completion. Additionally, when considering the issue of feasibility, qualitative information about the difficulties that arose in the completion of the trainer's report should also be considered essential.

Study methods

In this study pairs of trainers were asked to complete the proposed trainer's report form independently in relation to one trainee. The pilot study took place over a three-month period (October-December, 1995).

The report form

The report form used was that shown in appendix 5.2. It begins with detailed guidance about how items can be assessed and how the form should be completed. For each item the trainer is asked to indicate the type of assessment used and their overall judgement as to whether the trainee has reached the standard for independent general practice. Failure on a single item would result in failure of the whole report.

An example of the recording format is given overleaf.

Figure 4.4: Field testing study - example of recording format on draft trainer's report

2: the doctor is able to examine each system and each organ proficiently

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐NO☐

On the opposing page to the recording format details are provided on the standards to be used in reaching a judgement for each item (the standards being those developed in study four). An example is given below.

Figure 4.5: Field testing study - example of format for standards in draft trainer's report

2: the doctor is able to examine each system and each organ proficiently

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor is unable to undertake successfully a comprehensive examination or an important piece of examination	1,3

Sources other than trainer	*Specific methods for items marked 3*
partner, nurses, consultant. Diplomas may be taken into account.	OSCE, check list for each system/organ

Structure of study

All Directors of Postgraduate General Practice Education were asked to provide lists of practices in which there were two trainers; a list of 511 such practices was collected and all were invited by letter to be involved in the study. The letter explained that it was vital that both trainers and the trainee should be willing to be involved and that the study would require a substantial amount of time to be devoted to assessment, in acknowledgement of which a small financial payment would be made to those practices completing the study (the payment being £100 per practice). 69 pairs of trainers and their trainees initially indicated their agreement to be involved in the study.

One month before the beginning of the study period an information pack was sent to each of the participating practices. This pack contained information for the trainers which highlighted the purpose of the study, the need for independent assessments by each trainer, the need to avoid discussion about the outcomes of the assessment, and the need to assess trainees as though they were reaching the end of their training. A similar sheet was also supplied for the trainee describing the reasons for, and the process of, the study, and highlighting the anonymity of the report form and the opportunity for them to withdraw at any stage from the study. This sheet also indicated that discussion of the results with the trainers should be left until after the completion of the study in order to minimise the chances of discussion between the trainers during the study. The pack also contained two copies of the draft structured trainer's report, one for each trainer; other than being different colours (to allow separation of the "lead" trainer (i.e. the trainer with principal responsibility for the training) from the "other" trainer), the forms were identical. A reminder letter about completing the forms was sent at the half-way stage.

At the end of the study trainees were also sent a brief questionnaire asking about the difficulties that had arisen (appendix 4.6). This asked about the timing of the study relative to their training year, and the difficulties that arose as a result of the trainers in their practice completing the trainer's report form.

Calculation of results

Indicative discriminatory power was assessed by calculating, for each item, the proportion of trainees failed by one or both trainers.

Inter-rater reliability was assessed by calculating the proportion of pairs of trainers agreeing on the result of their assessment; the denominator for this calculation was the number of trainer pairs where a result was recorded by both trainers. This proportion was calculated for each item and also for the overall result. Where disagreement occurred between the trainers an analysis of bias was also undertaken (to see if there was systematic bias in the way the two trainers assessed the trainee) using McNemar's test (Bland, 1987). In addition for the overall result the *kappa* coefficient of agreement (Cohen, 1960), and its standard error (Streiner and Norman, 1995), were calculated. To test the possibility that the trainers had colluded in the use of the report form, for each item the proportion of instances in which both trainers recorded using the same combination of assessment methods was calculated (the denominator being the number of instances for which paired data were available).

To obtain information on the feasibility of the trainer's report form two approaches were used. Firstly numerical data was obtained by establishing the proportion of "lead" trainers who had not completed each item. Secondly qualitative data was obtained by asking all trainers to comment on the use of the report form under four headings

(appendix 4.7): "please list the difficulties that arose in completing this form"; "please list improvements that you would like to see in the report"; "please provide any hints for future users of the report"; and "any other comments".

At the end of the study practices were asked to return their report forms. A reminder was sent after two weeks and again, if no response, after a further four weeks at which time they were asked whether there was any particular reason for not completing the study.

4.3 Conclusions

Detailed methods for each of the five research studies have been developed and presented. In chapter five, the results of each of these studies are presented.

CHAPTER FIVE - RESULTS

This chapter presents the results of the five research studies. For each study the results are preceded by a brief summary of the aims of the study.

5.1 Study 1: Determining an appropriate structure for a trainer's report

5.1.1 Aims

The aim of this study is to determine an appropriate structure for the trainer's report. In particular, through a consideration of how a trainer's report would need to function, views on the structure of a new trainer's report that is most likely to be effective in fulfilling this function are sought. It is also intended to use this process of consultation with trainers to assess the degree to which the rest of the research programme is likely to answer the most important issues - that is, to test the validity of the research proposals.

5.1.2 Results

Thirteen groups agreed to be interviewed (seven in the Oxford region, six from the other four regions); two replied indicating that they did not wish to be interviewed, and two did not reply to either of the two letters. A list of the interviews undertaken is given in table 5.1.1 (overleaf). Thirteen interviews were undertaken over a period of eight months involving a total of 155 trainers (a mean of 11.9 trainers per group) which represents approximately 4.5% of all trainers in the United Kingdom at the time (information from all regions, July 1994). The interviews took a mean of 71.5 minutes.

Table 5.1.1: Details of interviews held with trainers' groups

Place	Date	Number of trainers	Duration (minutes)
Kettering	11.7.94	8	100
Slough	6.9.94	11	55
High Wycombe	12.9.94	13	70
Reading	13.9.94	22	70
Northampton	15.9.94	10	55
Aylesbury	30.9.94	11	70
Milton Keynes	2.11.94	7	60
Black Country 1	28.1.95	15	75
Black Country 2	28.1.95	13	70
Belfast	16.2.95	15	85
Lancaster	13.3.95	10	85
Borders (Melrose)	22.3.95	8	80
Preston	4.4.95	12	55

Outcome of interviews - report design

The detailed minutes of the interviews are contained in appendix 5.1.

Current systems - their availability, their pros and their cons:

All thirteen groups were currently undertaking some form of regular formative educational assessment. In a number of groups trainer-based assessment was supplemented by additional assessment involving assessors from outside of the practice (usually another trainer or the course organiser), some of which involved assessment solely of the trainee whilst others also included an assessment of the training. The assessment techniques used by trainers varied in their degree of formality, with some trainers undertaking an agreed formalised system (e.g. the West Midlands Region Formative Assessment Package or the North West Region Formative Assessment Package); a number of groups used the Manchester Rating Scales as part of the assessment. For many of the groups, the formative assessment package was not prescriptive.

The major perceived advantage of the systems currently in use were that the emphasis lay in formative assessment rather than summative assessment. In particular, many trainers expressed concerns about whether or not a trainer, as a result of summative assessment, should be put in the position of jeopardising an individual's career. The major perceived disadvantage of the current systems concerned the absence of any criterion referencing in the assessments made. A concern expressed by those groups with experience of rating scales was that such scales were difficult to use as formal assessment tools (although trainers frequently commented that, whilst there were too many scales to be used regularly for assessment, the headings were very useful as foci for

assessment). Other concerns expressed included that of the absence of any wide agreement on the contents of the assessment.

Four groups had considered the issue of a trainer's report for summative assessment in some detail. One group (Belfast) considered that the MRCGP examination provided a good form of summative assessment, with all trainees taking the examination. Two groups (Reading and Preston) had made a tentative exploration into the possibility of developing their own trainer's report, just prior to the interview. One group (Aylesbury) recognised that the use of structured references at the end of training was akin to a trainer's report, but that it lacked a uniform structure or a uniform set of standards against which trainees were assessed.

Suggestions for design of structured trainer's report:

From the thirteen groups four suggestions were made by at least six of the groups; no other suggestions were made by more than three groups. Ten groups wanted a simple answer format, not a rating scale; of the groups specifying a choice, 4 wished for a *"yes/no"* format, and 4 requested a *"yes/no/don't know"* format. Eight groups wanted input from others involved in training when the trainer's report was being completed; for some groups this was felt to be particularly important if a trainee was likely to fail the report, whilst for other groups this was judged particularly important if some clinical skills (e.g. intimate examinations) were to be assessed. Six groups specifically stated that they wished to see "criterion-referencing" rather than "peer-referencing". Six groups stated that the assessments should include clear records of the evidence on which any judgement was based.

One suggestion made by three groups was questioned by another of the groups: this concerned the possibility of using the results of formative assessments as the basis of the summative assessment, provided that failures were not "carried forward". The questioning arose because there was concern that formative assessment may not be adequately criterion-referenced, whereas summative assessment ought to be.

Outcome of interviews - concerns about previous trainees

In all groups at least one trainer had experienced concerns about at least one trainee. In most instances these were to do with attitudinal issues rather than knowledge, although in a number of cases the problems were more global in nature. In two instances the problems related to health.

When asked whether a structured trainer's report might have helped when dealing with these trainees, most groups who had time to consider this question felt that it would have helped. One group was unsure. In particular, it was felt that a trainer's report would have provided explicit standards against which the worrying trainee could have been assessed, that the report would have provided a structure for analysing exactly where the problems lay, and that a report form may also have provided a degree of protection against the fear that a trainee might make a legal challenge to the trainer's opinion.

5.1.3 Summary of findings

This study is based on interviews with 13 groups of trainers representing, in total, approximately 4.5% of trainers in the United Kingdom. The findings from these interviews are:

1. There is evidence that trainers have experienced problems with trainees.

2. These problems appear usually to be attitudinal in nature, are sometimes global and sometimes relate to health issues.
3. In the design of a new report the results of this study suggest that:

- there is a need for agreed content areas for the assessment;
- trainers want a trainer's report to be criterion-referenced;
- trainers would prefer the use of a simple record for the judgement made - either 'yes/no' or 'yes/no/don't know';
- trainers believe that it should be acceptable to include the views of others involved in training in their assessment;
- the report form should encourage the keeping of clear records of the evidence used in coming to a judgement;
- a report form may help them to focus on exactly where the problem lies and whether it is at a significant level.

4. The findings support the studies proposed for the development of this report form.

5.2 Study 2: Selecting appropriate contents

5.2.1 Aims

The aim of this study is to select appropriate content for a trainer's report as part of a summative assessment blueprint. To enable a comprehensive blueprint for the summative assessment process to be developed opinion is to be sought on the relative importance of attributes for independent practice; from those judged most important, the contents of a trainer's report are to be selected on the basis of the instruments judged most appropriate for the assessment of those attributes.

A secondary aim of this study is to obtain a numerical estimate of the frequency with which trainers have concerns about the performance of trainees.

5.2.2 Results

Response rates

Of the 1296 questionnaires sent to trainers 41 were returned because the recipient felt that they were ineligible for the study - 28 were no longer training, 9 had retired from general practice and 4 were on long-term absence from practice. Of the remaining 1255 eligible for inclusion 985 were returned of which 974 could be included in the study - an adjusted response rate of 77.6% (974/1255). This represents over one-quarter of the trainers in the United Kingdom (information from all regions, July 1994).

Characteristics of respondents

The mean age of respondents was 44.3 years; 867 (89.0%) were male. The mean total list size for their practices was 8806 patients, served by a mean of 4.81 partners including the trainer. They had been training for a mean of 7.8 years. Most of the respondents (72.6%) would usually have the trainee based in their practice for 9-12 months at a time.

Respondents were compared with non-respondents for gender and year of qualification (obtained from the Medical Register (General Medical Council, 1993)). Reliable information was available for 269 of the 281 non-respondents. There were no statistically significant differences between respondents and non-respondents for gender (107/974 female vs. 23/269 (χ^2 =1.34, df=1, p=0.24) or for year of qualification (mean 1973.1 (SD 6.88) vs. 1973.9 (SD 6.49) standard error of the difference = 0.45).

Responses to questions

Importance:

Table 5.2.1 (p.134) lists the proportion of respondents indicating an importance score of 4 or 5 for each item. Within the table the elements are listed in descending order of percentage of respondents indicating a high importance score, the items being separated into the five categories used in the questionnaire.

In the selection of contents based on the degree of importance of attributes, table 5.2.2 (p.136) demonstrates the effect of using different cut-off levels of importance score on the number of items in each category that would be included. It can be seen from table 5.2.2 that a cut-off level of at least 70% of respondents indicating an importance score of 4 or 5 is needed to ensure that at least one item from each of the five main categories would be included.

Assessment methods:

To provide information about the respondents' choice of assessment methods table 5.2.3 (p.137) indicates the following three measurements for each of those items for which 70% or more of respondents indicated an importance score of 4 or 5: those indicating the trainer's report alone; the modal response; and those indicating a response that did not include the trainer's report at all.

Table 5.2.1: Content study - proportion of respondents indicating an importance score of 4 or 5 on a 5-point scale

Element	% indicating importance score of 4 or 5 (95% CI)
1. PATIENT CARE	
The doctor can recognise common physical, psychological and social problems	98.0 (97.0-98.8)
The doctor diagnoses and manages acute emergency situations appropriately	96.1 (94.6-97.2)
The doctor responds appropriately to requests for urgent attendance at patients	91.3 (89.3-93.0)
The doctor is able to give an intravenous injection	90.3 (88.2-92.1)
The doctor demonstrates a broad knowledge of all aspects of the appropriate use of drugs (including actions, interactions, side effects, costs, and legal aspects)	89.4 (87.5-91.4)
The doctor is able to examine each system (e.g. cardiovascular, respiratory) and each organ (e.g. eye, ear) proficiently	89.2 (87.3-91.2)
The doctor is able to undertake basic cardio-pulmonary resuscitation	88.4 (86.4-90.4)
The doctor undertakes appropriate examination with appropriate consideration of the patients needs and feelings	88.0 (86.0-90.1)
The doctor is able to use the sphygmomanometer proficiently	87.1 (85.0-89.2)
The doctor is able to give an intramuscular injection	86.7 (84.6-88.9)
The doctor is able to use the vaginal speculum proficiently	86.2 (84.0-88.4)
The doctor is able to undertake a vaginal examination proficiently	85.6 (83.4-87.8)
The doctor is able to undertake a cervical smear proficiently	85.2 (82.9-87.4)
The doctor is able to use the stethoscope proficiently	84.8 (82.5-87.1)
The doctor is able to undertake a rectal examination proficiently	82.1 (79.7-84.5)
The doctor is able to use the peak flow meter proficiently	81.2 (78.7-83.7)
The doctor is able to use the auroscope proficiently	80.8 (78.3-83.3)
The doctor has the knowledge and skills to deal with life events and crises	79.8 (77.2-82.3)
The doctor is able to assess the mental state proficiently	78.7 (76.1-81.3)
The doctor chooses appropriate management for each problem with the patient (including the care of chronic problems)	76.8 (74.1-79.4)
The doctor provides appropriate care and support for patients and their families	76.3 (73.6-79.0)
Within the assessment the doctor includes patients' beliefs, ideas, concerns, effects and expectations	74.9 (72.2-77.6)
The doctor is able to use the ophthalmoscope proficiently	72.0 (69.2-74.9)
The doctor copes with the anxieties felt as a result of unstructured presentations, difficulty in reaching conclusions, and lack of continuous patient monitoring	67.4 (64.4-70.4)
The doctor makes effective use of the records	65.8 (62.8-68.8)
The doctor considers and follows up psychological and social factors	65.4 (62.4-68.4)
The doctor understands the importance of involving and educating patients	64.9 (61.9-67.9)
The doctor is able to use time as a diagnostic and therapeutic tool	64.4 (61.4-67.4)
The doctor has a knowledge of available agencies and resources and the skills to refer appropriately	62.3 (59.3-65.4)
The doctor uses management plans which include effective use of other members of the team	61.8 (58.7-64.9)
The doctor demonstrates an understanding of the effect of social and environmental circumstances on the patient	57.5 (54.4-60.6)
The doctor understands the principles involved in prevention in general practice (including case finding, screening, health education and monitoring of preventive activities)	56.3 (53.2-59.4)
The doctor understands the principles of problem definition (including the use of hypothesis formation and testing)	55.5 (52.3-58.6)
The doctor is able to provide effective preventive services to individual patients	53.4 (50.3-56.6)
The doctor is able to use the patellar hammer proficiently	49.1 (45.9-52.2)
The doctor has a knowledge of the systems used to identify individuals and sections of the practice population (e.g. disease registers, computerised registration data)	41.7 (38.6-44.9)
The doctor is aware of the costs of his/her activities and recognises the limits to those costs	37.3 (34.3-40.4)

The doctor is able to use the ECG proficiently	37.0 (34.0-40.1)
The doctor is able to provide effective preventive services to the population	34.4 (31.4-37.4)
The doctor is able to use the tuning fork proficiently	34.1 (31.1-37.1)
The doctor is able to use the proctoscope proficiently	33.1 (30.1-36.0)
The doctor is able to use the laryngoscope proficiently	7.4 (5.8-9.3)
2. COMMUNICATION	
The doctor demonstrates effective communication skills when dealing with patients	94.9 (93.3-96.2)
The doctor demonstrates understanding and respect for colleagues	68.7 (65.8-71.6)
The doctor uses his/her knowledge of the practice and of patients appropriately in various contacts (e.g. practice or team meetings, telephone contacts, contacts with families)	53.8 (50.7-57.0)
The doctor has an understanding of the importance of meetings and discussion with colleagues	52.0 (48.9-55.2)
The doctor demonstrates the skills to discover the strengths and weaknesses of colleagues and their need for support	38.3 (35.2-41.4)
3. ORGANISATION	
The doctor is aware of his/her own limitations, the skills of others and the ability to delegate appropriately	82.0 (79.5-84.3)
The doctor is able to manage his/her own time	81.0 (78.5-83.5)
The doctor understands his/her obligations according to the NHS contract and regulations	70.1 (67.2-72.9)
The doctor understands the importance of the need to manage a practice effectively	64.1 (61.0-67.1)
The doctor has a knowledge of the most important sections of the NHS contract and regulations with regard to sources of income and superannuation	47.2 (44.1-50.4)
The doctor is able to take appropriate action when organisational problems are identified	47.0 (43.9-50.2)
The doctor has a knowledge of the most important organisational aspects of practice and partnership (including agreements, accounts, building, tax)	44.5 (41.4-47.7)
The doctor is able to monitor aspects of practice activity	44.4 (41.2-47.5)
The doctor understands the principles of successful introduction of change and innovation	42.2 (39.1-45.3)
The doctor understands medico-social legislation and the impact of this on the patient	37.7 (34.6-40.7)
The doctor understands the application of new technology to general practice	31.8 (28.9-34.8)
The doctor knows how and where to intervene in the community on behalf of others	25.4 (22.7-28.2)
The doctor is able to determine and respond to the health needs of the community	24.5 (21.8-27.3)
The doctor has an understanding of the basic methods of research as applied to general practice	15.6 (13.3-17.9)
4. PROFESSIONAL VALUES	
The doctor possesses and applies ethical principles	80.5 (78.0-83.0)
The doctor is able to maintain his/her own physical and mental health	78.3 (75.7-80.9)
The doctor is willing to accept appropriate responsibility for patients, partners, colleagues and others	71.4 (68.5-74.2)
The doctor shows tolerance, respect and flexibility when responding to the ideas of others	67.0 (64.0-69.9)
The doctor is aware of the factors that influence the relationships between personal and professional life	65.8 (62.8-68.8)
The doctor is aware of his/her own values, beliefs and attitudes; how they are influenced; and how they affect others.	64.7 (61.7-67.7)
The doctor is willing to undergo peer review and is able to give and receive criticism	56.4 (53.3-59.5)
The doctor recognises the social cultural and organisational factors that define and affect his/her work	43.3 (40.2-46.5)
5. PERSONAL AND PROFESSIONAL GROWTH	
The doctor is able to identify strengths and weaknesses in his/her performance	76.5 (73.8-79.1)
The doctor is aware of the factors that limit his/her effectiveness	61.1 (58.1-64.2)
The doctor is able to manage and overcome the factors that limit his/her effectiveness	57.8 (54.6-60.9)
The doctor can define his/her own educational needs and appropriate methods of meeting those needs	56.1 (53.0-59.3)
The doctor can recognise, define and respond to change, including changing needs in patients, colleagues and the community	49.1 (46.0-52.3)
The doctor is able to produce change in self and others	41.3 (38.2-44.4)

Table 5.2.2: Content study - effect of different cut-off importance scores on the number of items to be included from each category (total number of possible items = 75)

Percentage indicating importance score of 4 or 5	Patient care (N = 42)		Communication (N = 5)		Organisation (N = 14)		Professional values (N = 8)		Personal and professional growth (N = 6)		Total number (N = 75)	
	n	%	n	%	n	%	n	%	n	%	n	%
20%	41	97.6	5	100.0	13	92.9	8	100.0	6	100.0	73	97.3
30%	41	97.6	5	100.0	11	78.6	8	100.0	6	100.0	71	94.7
40%	36	85.7	4	80.0	9	64.3	8	100.0	6	100.0	63	84.0
50%	34	81.0	4	80.0	4	28.6	7	87.5	4	66.7	53	70.7
60%	30	71.4	2	40.0	4	28.6	6	75.0	2	33.3	44	58.7
70%	23	54.8	1	20.0	3	21.4	3	37.5	1	16.7	31	41.3
80%	17	40.5	1	20.0	2	14.3	1	12.5	0	0.0	21	28.0
90%	4	9.5	1	20.0	0	0.0	0	0.0	0	0 0	5	6.7

Table 5.2.3: Content study - responses on favoured assessment methods for those elements for which 70% or more of respondents indicated an importance score of 4 or 5 (wr = written exam, ext = external observation, sub = trainee project, rpt = trainer's report, o = other).

Element	% favouring trainer's report alone	Modal response	% favouring methods of assessment not including trainer's report at all
1. PATIENT CARE			
The doctor can recognise common physical, psychological and social problems	9.9	wr,ext,rpt	13.7
The doctor diagnoses and manages acute emergency situations appropriately	20.6	wr,ext,rpt	8.3
The doctor responds appropriately to requests for urgent attendance at patients	52.0	rpt	5.4
The doctor is able to give an intravenous injection	48.7	rpt	21.0
The doctor demonstrates a broad knowledge of all aspects of the appropriate use of drugs (including actions, interactions, side effects, costs, and legal aspects)	5.5	wr,rpt	24.4
The doctor is able to examine each system (e.g. cardiovascular, respiratory) and each organ (e.g. eye, ear) proficiently	17.9	ext,rpt	17.6
The doctor is able to undertake basic cardio-pulmonary resuscitation	15.5	ext	43.2
The doctor undertakes appropriate examination with appropriate consideration of the patients needs and feelings	14.3	ext,rpt	15.6
The doctor is able to use the sphygmomanometer proficiently	24.8	ext,rpt	25.7
The doctor is able to give an intramuscular injection	48.6	rpt	20.5
The doctor is able to use the vaginal speculum proficiently	43.2	rpt	20.2
The doctor is able to undertake a vaginal examination proficiently	41.8	rpt	20.0
The doctor is able to undertake a cervical smear proficiently	40.6	rpt	20.1
The doctor is able to use the stethoscope proficiently	25.5	ext,rpt	24.7
The doctor is able to undertake a rectal examination proficiently	44.4	rpt	21.1
The doctor is able to use the peak flow meter proficiently	27.8	ext,rpt	23.7
The doctor is able to use the auroscope proficiently	23.7	ext,rpt	25.0
The doctor has the knowledge and skills to deal with life events and crises	25.2	rpt	8.8
The doctor is able to assess the mental state proficiently	15.3	ext,rpt	18.2
The doctor chooses appropriate management for each problem with the patient (including the care of chronic problems)	12.4	wr,ext,rpt	14.8
The doctor provides appropriate care and support for patients and their families	39.7	rpt	5.9
Within the assessment the doctor includes patients' beliefs, ideas, concerns, effects and expectations	10.0	ext,rpt	16.0
The doctor is able to use the ophthalmoscope proficiently	24.1	ext,rpt	25.8
2. COMMUNICATION			
The doctor demonstrates effective communication skills when dealing with patients	6.6	ext,rpt	14.4
3. ORGANISATION			
The doctor is aware of his/her own limitations, the skills of others and the ability to delegate appropriately	39.4	rpt	4.9
The doctor is able to manage his/her own time	47.0	rpt	4.4
The doctor understands his/her obligations according to the NHS contract and regulations	10.2	wr,rpt	30.2
4. PROFESSIONAL VALUES			
The doctor possesses and applies ethical principles	31.3	rpt	8.6
The doctor is able to maintain his/her own physical and mental health	64.6	rpt	7.7
The doctor is willing to accept appropriate responsibility for patients, partners, colleagues and others	62.3	rpt	4.5
5. PERSONAL AND PROFESSIONAL GROWTH			
The doctor is able to identify strengths and weaknesses in his/her performance	38.6	rpt	7.5

Table 5.2.3 demonstrates a number of findings. Firstly, if the modal responses alone are considered (the third column), for only one of the 31 attributes does the modal response not include the trainer's report (the element concerning the ability to undertake basic cardio-pulmonary resuscitation); all of the other modal responses include assessment by the trainer's report - for five attributes the modal response also includes assessment by a written exam, and for 13 assessment by an external observation of performance. This, as might be expected, is supported by the fourth column which demonstrates that when the percentage favouring methods of assessment that did not include a trainer's report at all is considered, this same element is separated from all the other elements by a considerable margin. The finding that is perhaps less predictable is that there is no clear relationship between the percentage of trainers favouring assessment by a trainer's report alone (the second column) and the percentage favouring methods that did not include the trainer's report (the fourth column) - for example of those elements for which >40% of respondents favoured a trainer's report alone, the results in the fourth column vary from 4.4 to 21.1%. This may be explained, at least in part, by the finding that most of the elements for which >20% of respondents favoured methods that did not include the trainer's report were elements describing clinical skills - this may reflect the hope of respondents that the chosen method of external observation would allow assessment of clinical skills rather than the need to assess these skills by means of a trainer's report.

Other results:

No additional items for inclusion in the trainer's report were suggested by more than 5% of the respondents.

There were 952 responses to the question "have you ever considered not signing up a trainee on form VTR1?". 251 replied that they had (26.4% (SE 1.43%)). Based on a mean of 7.8 years of training, this represents 3.4% of trainers considering not signing up a trainee on form VTR1 each year; the corollary of this is that an individual trainer would, on average, consider not signing the VTR1 form once every 29.5 training years

5.2.3 Summary of findings

As a result of a good response rate this study involved over one-quarter of all trainers in the U.K. The main results are that:

1. Problems with trainees appear to occur with a relatively high frequency (approximately one in thirty trainees per year).
2. Trainers are able to discriminate between attributes both on the basis of importance and on the basis of methods of assessment; consequently their views can be used to develop an assessment blueprint.
3. If a balance between feasibility and inclusivity is to be achieved, a cut-off of 70% of respondents indicating an importance score of 4 or 5 on a five-point scale ensures that at least one item from each category is included; this cut-off would result in a report containing 31 items.

4. In completing the blueprint, of the 31 items chosen on the basis of this cut-off for importance, only one was consistently excluded from assessment by the trainer's report; for five of these 31 the modal response also included assessment by a written exam, and for 13 assessment by an external observation of performance.
5. No additional attributes were suggested by more than 5% of trainers.

5.3 Study 3: Assessing content validity

5.3.1 Aims

The aim of this study is to assess the views of another group of key players on the validity of the content of the proposed trainer's report.

5.3.2 Results

Respondents

Three questionnaires were returned by the Post Office leaving 217 potential responders. 159 completed questionnaires were received, resulting in an adjusted response rate of 73.3% (159/217). The mean age of respondents was 31.1 years.

Information on gender was available for 158 respondents and all 61 non-respondents. There was no significant difference between respondents and non-respondents (women representing 87 of 158 respondents and 32 of 61 non-respondents (chi-square=0.12, df=1, p=0.8)).

Responses

The percentage of respondents agreeing that the item was needed in general practice (i.e. responded "agree" or "strongly agree") is shown in table 5.3.1 (p.143). The first column contains the items; the second column contains the percentage of respondents who agreed; the third column contains the 95% confidence interval for the percentages (Gardner et al. 1992).

The responses to the questions about whether it is reasonable to assess the item by means of a trainer's report are presented in table 5.3.2 in three ways (p.145). The first column contains the items (in abbreviated form); the second column indicates the percentage who agreed ("agree" or "strongly agree"); the third column shows the percentage who neither agreed nor disagreed, and the fourth column the percentage who disagreed ("disagree" or "strongly disagree"). The fifth column indicates the difference between the percentages who agreed and disagreed, and the sixth column contains the 95% confidence intervals for these differences (Gardner et al. 1992).

The denominators used in tables 5.3.1 and 5.3.2 to calculate the percentages were the total number of responses to that question which varied between 156 and 159. Table 5.3.1 is presented in decreasing order of agreement; table 5.3.2 uses the same order as table 5.3.1.

Table 5.3.1 demonstrates that for 31 of the 33 attributes considered more than 85% of respondents agreed that the item was a skill, attitude or piece of knowledge that was

needed in general practice. The two exceptions were those two items included in the questionnaire which had already been excluded from the trainer's report based on the survey of trainers (study two) and included in this study purely as a test of acquiescence bias.

Table 5.3.1: Content validity study - percentage of respondents indicating agreement with the statements for each item

Item	Percentage indicating that item needed in general practice	
	%	95% CI
The doctor is aware of his/her own limitations, the skills of others, and the ability to refer or delegate appropriately	99.4	96.6-100.0
The doctor demonstrates effective communication skills when dealing with patients	99.4	96.6-100.0
The doctor responds appropriately to requests for urgent attendance at patients	99.4	96.6-100.0
The doctor diagnoses and manages acute emergency situations appropriately	99.4	96.6-100.0
The doctor can recognise common physical psychological and social problems	98.7	95.5-99.9
The doctor undertakes appropriate examination (including investigations)	98.7	95.5-99.9
The doctor undertakes examination with appropriate consideration of the patient's needs and feelings	98.7	95.5-99.9
The doctor can use the sphygmomanometer proficiently and interpret the findings made	98.7	95.5-99.9
The doctor can use the auroscope proficiently and interpret the findings made	98.7	95.5-99.9
The doctor can undertake the cervical smear and interpret the findings made	98.7	95.5-99.9
The doctor can use the stethoscope proficiently and interpret the findings made	98.7	95.5-99.9
The doctor can use the vaginal speculum proficiently and interpret the findings made	98.7	95 5-99.9
The doctor provides appropriate care and support for patients and their families	98.1	94.6-99.6
The doctor is able to examine each system and each organ proficiently	98.1	94.6-99.6
The doctor is able to give an intramuscular or subcutaneous injection proficiently	98.1	94.6-99.6
The doctor is able to identify strengths and weaknesses in his/her performance	97.5	93.7-99.3
The doctor is able to manage his/her own time	97.5	93.6-99.3
The doctor chooses appropriate management for each problem with the patient (including the care of chronic problems)	96.9	92.8-99.0
The doctor can use the peak flow meter proficiently and interpret the findings made	96.9	92.8-99.0
The doctor is able to examine the mental state proficiently and interpret the findings made	96.9	92.8-99.0
The doctor can undertake the vaginal examination proficiently and interpret the findings made	96.9	92.8-99.0
The doctor demonstrates a broad knowledge of all aspects of the appropriate use of drugs (including actions, interactions, side effects, costs and legal aspects)	96.2	92.0-98.6
The doctor has the knowledge and skills to deal with life events and crises	95.6	91.1-98.2
The doctor understands the obligations of a general practitioner according to the NHS contract and regulations	95.0	90.3-97.8
The doctor possesses and applies ethical principles	95.0	90.3-97.8
The doctor can undertake the rectal examination proficiently and interpret the findings made (does not include proctoscopy)	95.0	90.3-97.8

Table 5.3.1 continued:

The doctor can use the ophthalmoscope proficiently and interpret the findings made	93.1	87.9-96.5
The doctor is able to give an intravenous injection proficiently	93.1	88.0-96.5
The doctor is able to maintain his/her own physical and mental health to a level which enables him/her to discharge the duties of a general medical practitioner	91.8	86.4-95.6
The doctor is willing to accept appropriate responsibility for patients, partners, colleagues and others	90.6	84.9-94.6
Within his/her assessment the doctor includes the patients' beliefs, ideas, concerns, expectations and the effects of the problem	89.9	84.1-94.1
The doctor has an understanding of the basic methods of research as applied to general practice	60.7	52.7-68.4
The doctor can use the laryngoscope proficiently and interpret the findings made	15.7	10.4-22.3

Table 5.3.2: Content validity study - percentage of respondents agreeing, disagreeing or neither with assessment of items by means of a trainer's report

Item	% agree	% neither agree nor disagree	% disagree	difference between % agree and % disagree	95% CI of difference
Aware of his/her limitations	78.0	12.6	9.4	68.6	60.7-76.4
Communication skills when dealing with patients	78.5	15.2	6.3	72.2	64 7-79 6
Responds to requests for urgent attendance	69.2	17.6	13.2	56.0	47 1-64.9
Manages acute emergency situations appropriately	67.7	15.2	17.1	50.6	41.3-60.0
Recognises common problems	86.0	7.0	7.0	79.0	72.2-85 7
Undertakes appropriate examination	75.5	14.5	10.1	65.4	57.2-73.6
Undertakes examination with consideration	74.0	15.8	10.1	63.9	55.6-72.2
Can use the sphygmomanometer	59.6	17.7	22.8	36.7	26.6-46.8
Can use the auroscope	57.6	20.9	21.5	36.1	26.1-46.1
Can undertake the cervical smear	52.2	25.2	22.6	29.6	19.4-39.7
Can use the stethoscope	51.6	22.6	25.8	25.8	15.5-36.1
Can use the vaginal speculum	*42.8*	*23.9*	*33.3*	*9.4*	*-1.2-20.1*
Provides care and support for patients and families	69.8	22.0	8.2	61.6	53.3-69.9
Able to examine each system/organ	62.6	22.2	15.2	47.5	38.1-56.9
Able to give im or sc injection	43.0	28.2	28.8	14.1	3.6-24.6
Able to identify strengths/ weaknesses	72.4	18.9	8.8	63.5	55.3-71.8
Able to manage own time	62.0	22.2	15.8	46.2	36.7-55 7
Chooses appropriate management	78.5	15.8	5.7	72.8	65.4-80.1
Can use the peak flow meter	61.7	16.4	22.0	39.6	29.7-49.6
Able to examine the mental state	57.6	25.9	16.5	41.1	31.5-50.8
Can undertake the vaginal examination	*41.5*	*23.3*	*35.2*	*6.3*	*-4.4-17.0*
Demonstrates knowledge use of drugs	61.6	19.5	18.9	42.8	33.1-52.5
Can deal with life events and crises	70.5	17.0	12.6	57.9	49.1-66.6
Understands the obligations of a GP	62.3	24.5	13.2	49.1	39.9-58.2
Possesses and applies ethical principles	57.6	20.9	21.5	36.1	26.1-46.1
Can undertake the rectal examination	*38.4*	*26.4*	*35.2*	*3.1*	*-7.5-13.7*
Can use the ophthalmoscope	45.6	24.7	29.7	15.8	5.3-26.4
Able to give an iv injection	*37.1*	*27.7*	*35.2*	*1.9*	*-8.7-12.4*
Able to maintain physical and mental health	*37.7*	*30.8*	*31.4*	*6.3*	*-4.1-16.7*
Willing to accept appropriate responsibility	48.5	34.0	17.6	30.8	21.1-40.6
Includes the patients' concerns	70.1	20.4	9.6	60.5	52.0-69.0
Understands basic methods of research	*27.3*	*50.6*	*22.2*	*5.1*	*-4.4-14.6*
Can use the laryngoscope	8.2	24.8	66.0	-57.9	-66.4- -49.4

In table 5.3.2 it can be seen that for a sizable proportion of items there was a substantial group who neither agreed nor disagreed with assessment by means of a trainer's report; consequently a simple interpretation of the results based on the proportion agreeing or disagreeing would not present a full picture. The result considered to give the most information (because it removes the group who neither agreed nor disagreed) is the difference between the proportion agreeing and the proportion disagreeing with assessment by a trainer's report. This demonstrates that for 26 of the items, significantly more of respondents agreed that it was reasonable that the item was assessed by means of a trainer's report than disagreed. For 6 items (in italics) there was no significant difference whilst for one item (underlined) significantly more disagreed than agreed. For two items ("use of the ophthalmoscope" and "accepting appropriate responsibility for patients, partners, colleagues and others"), whilst significantly more agreed than disagreed, the proportion agreeing was less than 50%. This occurred because a large proportion (24.7% and 34.0% respectively) neither agreed nor disagreed with assessment of these items by means of a trainer's report.

Freetext comments

87 respondents made freetext comments. Comments made by two or more respondents are summarised in table 5.3.3 (below).

Table 5.3.3: Content validity study - summary of freetext comments made by respondents

Comment	Number
The report may be affected by the relationship between the trainer and trainee and consequently should be as objective as possible	26
Some of the skills may be best assessed by someone other than the trainer	16
The trainer will need to undertake an adequate amount of observation of an appropriate type	11
The standards set will need to be appropriate for general practice, being set neither too high nor too low	9
There will need to be adequate assessment of the trainers themselves	6
The report needs to have inter-rater reliability	6
It is important that the trainer's report is only one part of the assessment	5
There is a need for more assessment by the trainer than is currently undertaken	4
It is important that there is input from someone other than the trainer if difficulties arise	3
The report needs to be simple	2
The trainer's report should form part of an overall appraisal of the trainee	2
I am worried about the inclusion of a trainer's report in summative assessment	2

5.3.3 Summary of findings

Again the response rate to the questionnaire in this study was over 70%. In this study the main findings are that:

1. More than 85% of respondents agreed that the items selected in study two were important for independent general practice.
2. The low level of agreement for the two items that are not intended for inclusion in the trainer's report confirm that the high levels of agreement for the other items are not likely to have occurred as a result of acquiescence bias alone.
3. Although the results of the component of the study dealing with assessment by the trainer's report can be interpreted in different ways depending on the exact results

analysed, for 26 items more agreed with assessment by the trainer's report than disagreed.

4. A number of issues were highlighted by respondents. These were: that, because of the potential effect of the trainer-trainee relationship on the interpretations made, the criteria should be as objective as possible; that respondents supported the use of assessment by colleagues other than the trainer; that respondents recognised the need for close observation if the report was to be successfully completed; and that there was a need for appropriate (minimum) standards to be developed.

5.4 Study 4: Setting standards

5.4.1 Aims

The aim of this study is to develop minimum standards for each of the items selected for inclusion in the trainer's report. In addition, this study also considers how performance can be tested against these standards, and who is in the best position to observe it.

5.4.2 Results

Consensus conference

By the end of the conference conclusions from three groups were available for all thirty items considered. For all items it was possible to condense the conclusions of the groups into a maximum of four standards. These standards, along with the suggestions made for methods for collecting acceptable evidence and personnel whose views could be used, were used to form a draft trainer's report which was then submitted to the consultation

phase. This draft also contained a first draft of general guidance for trainers completing the report.

Consultation phase

46 replies (82.1%) were received; all of the 30 delegates from the consensus meeting replied whilst 16 of the 26 experts replied (61.5%). 33 of the 46 replies (71.7%) contained at least three comments.

Alterations to the draft were included if a change was suggested consistently in three or more replies, or a change was suggested that made the standards consistent with other standards without fundamentally altering the original standard, or a change was suggested that made the standard clearer without fundamentally altering the original standard. Alterations were excluded if the change suggested was one for which there was evidence that the consensus view would not have supported such a change.

The consultation phase resulted in no major modifications to the standards; a number of minor alterations to the standards or to the suggestions for methods of collecting evidence were included. The only major modification proposed was to divide one item ('the doctor undertakes appropriate examination with appropriate consideration of the patients needs and feelings') into two ('the doctor undertakes examination with appropriate consideration of the patients' needs and feelings' and 'the doctor undertakes appropriate examination (including investigations)'); this enabled clarification of the standards set. The resulting version of the report consequently contains 31 items.

A number of respondents commented on the use of the words *"persistently"*, *"repeatedly"* and *"appropriate"* within the standards. They commented that this left a significant amount to the judgement of the trainer and suggested that clarification about the meanings of these terms should be made in the introduction to the trainer's report.

Outcome

In total 79 standards were developed (a mean of 2.5 standards per item). For each of the proposed standards guidance was given as to the most appropriate methods for collecting evidence. The proposed methods fell into three categories - direct observation of the trainee by the trainer, tutorial-based discussion, and methods specific to the particular item under consideration. Table 5.4.1 (p.151) demonstrates the number of standards assessible in these three different ways. It can be seen that for all but one of the proposed standards (one of the standards dealing with the possession and application of ethical principles) delegates believed that direct observation was appropriate. In addition, guidance was also given about personnel whose evidence was thought to be acceptable for assessment purposes. Table 5.4.2 (p.152) lists the acceptable sources suggested and the number of items (out of the final total of 31) for which each is appropriate.

A number of suggestions were made about the layout. In particular, representatives of the JCPTGP were keen that, to encourage trainers to ensure that their assessment results were based on the standards set, the standards and the results pages should form a single

document, with the standards (along with the advice on the sources and assessment methods) being on the page facing the assessment results.

Table 5.4.1: Standards study - suggested methods of assessment for the 79 proposed standards

Method	*Examples*	*Number of standards with this method suggested (%)*
1. Direct observation	Joint consultation Video-taped consultations	78 (98.7)
2. Tutorial-based discussion	Random case analysis Case discussion	40 (50.6)
3. Specific methods	Notes review Complaints	61 (77.2)

Table 5.4.2: Standards study - acceptable sources of evidence for assessment for the 31 proposed items

Source of evidence	Number of items (%)	
Trainer	31	(100.0)
Partner	31	(100.0)
Consultant	17	(54.8)
Any primary health care team member	16	(51.6)
Nurse	10	(32.2)
Course organiser	4	(12.9)
Family planning clinic trainer	3	(9.7)
Patients	2	(6.5)
Diplomas	2	(6.5)
Pharmacist, Family Health Services Authority pharmaceutical adviser, practice manager, police, courts, previous employer, community psychiatric nurse, trained counsellor	1 each	(3.2)

Final draft report

The final version of the draft trainer's report, based on the outcome of both the consensus conference and the consultation phase, is reproduced as appendix 5.2.

5.4.3 Summary of findings

This study had two phases. It was apparent that, particularly for those involved in the consensus conference, there were high levels of commitment to the process. The main findings of this study are:

1. Standards have been set for all the items selected in study two and verified in study three.
2. The consultation phase resulted in few alterations to the recorded deliberations of the consensus conference.
3. Trainers will need to exercise judgement in their interpretation of the standards.
4. For all items (and all bar one individual standard) it should be possible to collect evidence by direct observation, thereby enhancing the overall validity of the report;
5. For all items (and all individual standards) personnel other than the trainer are considered to be acceptable sources of evidence for assessment.

5.5 Study 5: Assessing overall validity, inter-rater reliability and feasibility

5.5.1 Aims

The aim of this study is to assess the overall validity (by means of discriminatory power), inter-rater reliability and feasibility of the report form. The report form tested is that shown in appendix 5.2.

5.5.2 Results

Respondents

69 practices initially expressed interest in the study, but six withdrew because of unforeseen problems within the practice (five because of problems within the practice preventing them being able to concentrate on this study and one because the trainee no longer wished to be involved). From the remaining 63 practices, report forms were received from 52 pairs of 'lead' and 'other' trainers; single report forms were received from a further four practices (three from 'lead' trainers, and one from an 'other' trainer). Responses were therefore received from 56 practices (88.8% of 63). Of the 7 practices from which no forms were received two indicated that they had been unable to complete the study because of unexpected doctor absences but no information was available from the remaining five.

For the quantitative data analysis the 52 pairs of report forms were used (48 of which had corresponding trainee data available). For qualitative data all 108 report forms were used. Respondents came from 17 different regions of the United Kingdom. The degree to which the report forms were completed is shown in table 5.5.1 (overleaf). This demonstrates that no report form had fewer than 23 items recorded as having been assessed, with the vast majority having all items considered. Because it was recognised that the three-month duration of the study might prevent all items from being adequately assessed, when analysing the overall result a registrar was assessed as having passed if all completed items were passed and to be failed if one or more items had been failed.

Table 5.5.1: Field testing study - degree to which report forms were completed

no. of items completed (of 31)	lead trainer	other trainer
all	39	34
30	7	8
29	1	0
28	3	1
27	1	2
26	0	2
25	0	2
24	1	2
23	0	1
less than 23	0	0
total	52	52

Reliability

The results of the reliability testing are shown in the second column of table 5.5.2 (p.159). This shows that for all of the 31 items more than 90% of pairs of trainers agreed on their assessment. To consider whether there was a systematic trend in assessment (that is, whether the lead trainers as a group, or the other trainers as a group, were systematically more or less likely to fail the trainee) a McNemar's test (Yates correction) was undertaken. Based on a significant result being one of $p<0.05$, there was no evidence of significant systematic bias for any of the 31 items. To provide some information on whether or not it was likely that there had been collusion between the two trainers, a review of the assessment methods indicated by the two trainers was made. For all bar one item, less than half of the pairs of trainers indicated using the same combination of assessment methods. The one exception was the assessment of the use of the auroscope for which 56.5% (26 of 46 pairs) indicated using the same method (direct observation of the trainee by the trainer).

For the report form as a whole, eight of the 52 trainees would have been failed, 3 being failed by both trainers and 5 by one trainer only (three by the lead trainer and two by the other trainer). This results in a kappa coefficient for the overall result of 0.49 which indicates agreement of moderate strength (Brennan and Silman, 1992), with a standard error of 0.22. Information was available about six of these eight trainees which indicated that all six were in the first six months of their training at the time of this study.

Indicative discriminatory power

These results are presented in the second results column of table 5.5.2 (p.159). Nineteen of the 31 items would have resulted in at least one trainee being failed.

Feasibility

The final column of table 5.5.2 (p.159) indicates the proportion of the 52 principal trainers who had not completed the assessment for that item. This shows that for no item did more than 6% of the principal trainers fail to indicate that the item had been assessed. The greatest difficulty occurred with the assessment of the rectal examination and smear-taking.

The three freetext comments made most frequently by trainers in response to the three principal questions asked are listed in table 5.5.3 (p.160), along with the number of times the comment was made. This demonstrates that the major difficulties centred on the problems associated with the observation of some clinical skills, in particular intimate examinations (rectal and gynaecological examinations) and injections. This is supported by a number of trainers suggesting that this might be best done in other ways (in particular assessment in hospital posts prior to entering general practice, or by the use of mannequins). The guidance notes were not easy to use and a number of changes were suggested (in particular the provision of examples for how the completed report might look).

Trainers were also offered the opportunity to make any other comments they wished. Nine trainers made broadly negative comments about the report form, twelve commented that they could see it being useful when a trainee was poor but less so when the trainee was good; twelve commented that they found the report form useful.

Table 5.5.2: Field testing study - results on inter-rater reliability, relative likelihood of failure and feasibility for the items in the draft trainer's report

Item	% of trainer pairs agreeing [α]	% of trainees failed [α,β]	% of reports with no principal trainer record [γ]
1. PATIENT CARE			
a) The doctor can recognise common physical, psychological and social problems	94.2	5.8	0.0
b) The doctor diagnoses and manages acute emergency situations appropriately	98.0	2.0	0.0
c) The doctor responds appropriately to requests for urgent attendance at patients	98.0	2.0	1.9
d) The doctor is able to give an intravenous injection	100.0	0.0	3.8
e) The doctor demonstrates a broad knowledge of all aspects of the appropriate use of drugs (including actions, interactions, side effects, costs, and legal aspects)	92.2	9.8	0.0
f) The doctor is able to examine each system and each organ proficiently	98.0	4.0	3.8
g) The doctor undertakes appropriate examination (including investigations)	96.1	3.9	1.9
h) The doctor undertakes examination with appropriate consideration of the patients needs and feelings	100.0	0.0	1.9
i) The doctor is able to use the sphygmomanometer proficiently	100.0	0.0	0.0
j) The doctor is able to give an intramuscular injection	100.0	0.0	0.0
k) The doctor is able to use the vaginal speculum proficiently	100.0	2.3	3.8
l) The doctor is able to undertake a vaginal examination proficiently	100.0	2.3	3.8
m) The doctor is able to undertake a cervical smear proficiently	97.5	2.5	5.8
n) The doctor is able to use the stethoscope proficiently	100.0	0.0	0.0
o) The doctor is able to undertake a rectal examination proficiently	100.0	0.0	5.8
p) The doctor is able to use the peak flow meter proficiently	100.0	0.0	0.0
q) The doctor is able to use the auroscope proficiently	100.0	0.0	0.0
r) The doctor has the knowledge and skills to deal with life events and crises	96.1	3.9	0.0
s) The doctor is able to assess the mental state proficiently	100.0	0.0	0.0
t) The doctor chooses appropriate management for each problem with the patient (including the care of chronic problems)	92.0	10.0	1.9
u) The doctor provides appropriate care and support for patients and their families	100.0	0.0	3.8
v) Within the assessment the doctor includes patients' beliefs, ideas, concerns, effects and expectations	94.2	5.8	0.0
w) The doctor is able to use the ophthalmoscope proficiently	100.0	2.2	3.8
2. COMMUNICATION			
a) The doctor demonstrates effective communication skills when dealing with patients	100.0	2.0	1.9
3. ORGANISATION			
a) The doctor is aware of his/her own limitations, the skills of others and the ability to delegate appropriately	98.0	4.0	1.9
b) The doctor is able to manage his/her own time	91.8	10.2	3.8
c) The doctor understands his/her obligations according to the NHS contract and regulations	95.9	6.1	3.8
4. PROFESSIONAL VALUES			
a) The doctor possesses and applies ethical principles	100.0	0.0	0.0
b) The doctor is able to maintain his/her own physical and mental health	100.0	0.0	1.9
c) The doctor is willing to accept appropriate responsibility for patients, partners, colleagues and others	98.1	1.9	0.0
5. PERSONAL AND PROFESSIONAL GROWTH			
a) The doctor is able to identify strengths and weaknesses in his/her performance	98.0	3.9	1.9

α denominator is number of trainer pairs where a result was recorded by both trainers

β failed by one or both trainers

γ denominator is 52 principal trainers

Table 5.5.3: Field testing study - the three freetext comments made most frequently by trainers in the three main categories

Comment	No. of times reported
Difficulties experienced:	
Problems with undertaking assessment of clinical skills	22
Problems with understanding the guidance notes	15
Problems with recalling dates on which the assessments had been made	13
Suggestions for improvements to the report:	
Make changes to the guidance notes	24
Make alterations to the layout of the report	19
Make alternative arrangements for the assessment of clinical skills	8
Hints for future users of the report:	
Start using the report early in the training year	18
Keep written records of assessments (including dates)	13
Complete the report as you go along	7

Trainee feedback

Of the 48 trainees who completed the questionnaire at the end of the study 13 (27.1%) described difficulties that had arisen as a result of using the report. Four comments were made by more than one trainee: that problems arose in the assessment of intimate clinical examinations (4 respondents); that it appeared to take time to undertake the observations (3), which may have interfered with teaching time (2); and that it was insulting to have examination skills assessed at this late stage of training (2).

5.5.3 Summary of findings

Bearing in mind the intensity of the requirements of this study on trainers and trainees there were high levels of completion of report forms. The main findings are that:

1. There is evidence to suggest that the report form has inter-rater reliability, with over 90% agreement on the results within trainer pairs for all items, and a kappa coefficient overall of 0.49.
2. There is evidence that it is likely that the report form will have overall validity with 19 of the 31 items resulting in failure of at least one trainee of the sample considered.
3. There is evidence of feasibility with, for each item, less than 6% of trainers failing to indicate that an assessment had been completed.
4. Problems do exist. For trainers problems exist with the assessment of clinical skills, the guidance notes, keeping adequate records, and the layout of the form; for trainees problems exist with the assessment of clinical skills, and time for the assessment.
5. For future users this group of trainers suggested that the report form should be used from early in the training year, and that the assessments and the records should be maintained as they were completed.

5.6 Conclusions

This sequential series of five research studies has provided evidence on which a trainer's report for use by general practitioner trainers in the assessment of general practitioner trainees could be based; using the approach of a blueprint the content of the report form has been selected and standards (with suggested methods and personnel for the collection of evidence) developed. Evidence has also been provided which supports its content validity, overall validity, inter-rater reliability and feasibility.

In chapter six these results are now considered in detail. By assessing the limitations of the research studies, the weight that can be placed on the results presented in this chapter is considered. The results are also considered in the context of other published work and conclusions are drawn.

CHAPTER SIX - CONCLUSIONS FROM THE RESEARCH

In this chapter the results of the programme of research studies are scrutinised. To enable conclusions to be drawn for each of the four areas considered in the research (report structure (study one), content (studies two and three), standards (study four) and field testing (study five)) four questions are asked. Firstly, what are the main results of the research? Secondly, what methodological issues have arisen in the research, and how does this affect the weight that can be placed on the results found? Thirdly, what are the principal issues that arise from these weighted results? Fourthly, how do these results sit within the context of other published work? Conclusions from the research will then be drawn.

6.1 The structure of a trainer's report (study one)

6.1.1 Main results

The principal results of study one are: that the report should use an answer format that is simple; that trainers should be able to involve others involved in training when completing the report; and that the report should allow documentation of the evidence used to support the judgements made. Study one also demonstrated that trainers wished for the content of a report to be defined and for criterion-based standards to be developed (thereby supporting the subsequent research studies). The study also demonstrated that the development of a summative assessment process for general practice is supported by the finding that trainers have experienced problems with trainees; particular support for the place of a trainer's report within a summative assessment process has been provided by the finding that these problems appear usually to be attitudinal in nature, are sometimes global and sometimes relate to health issues; it

is unlikely that any of the three other proposed assessment instruments would provide evidence about performance in these domains.

6.1.2 Methodological issues

The group interview did enable information to be obtained from a large number of people in all the areas in which it was sought. Although the study was not designed to allow any comparison with results that might have been established from individual interviews with a similar number of trainers the results do support the view that a group interview is a valuable method for the exploration of ideas from large numbers of people (Frey and Fontana, 1991). In particular the establishment of a group response does reduce considerably the onus on the interviewer in the interpretation of the data collected.

However, the problem of the interviewer-effect does remain (Frey and Fontana, 1991), and the extent of this problem in this study remains unknown; whilst the use of a single interviewer, with some small group skills, might ensure a reasonably uniform approach to the interview and a reasonable degree of control over the interview to maintain the discussion around the intended focus, examination of the minutes does demonstrate that the answers to some questions were either sought or recorded in greater depth as the interviewer moved towards the later interviews (for example, as evidenced by the more specific records of the numbers of trainers describing previous concerns with trainees) and that there is more consistency in the nature of the recorded responses in the later interviews. Whilst the former may well be an interviewer-effect it is not clear whether the latter is also an interviewer-effect or whether, because the interviews were spread over eight months, this resulted from a gradual increase in the knowledge of trainers as

to what might be expected of them (and more time to consider their views) with a resultant trend towards more consistent responses. The absence of either a second interviewer or tape recordings of the interviews to corroborate the results does mean that it is not possible to be sure that this did not occur and should be considered a methodological oversight. Consequently, it is important that no particular weighting is given to responses given in later interviews.

The other methodological concern centres on the distribution of the groups interviewed, with a particular preponderance of groups from one region. Examination of the minutes, demonstrates that there is no pattern in which the responses from the groups from the Oxford region are consistently different from those of the other groups. The results of this study are also broadly in line with those from the work undertaken in both the North Thames (West) Region (Rhodes and Styles, 1995) and in the Trent Region (Hasler, 1994), suggesting that, despite the geographical limitations, the results of this study can be considered to be generalisable.

Because the selection of trainers' groups was not systematic (and may not therefore be representative) this study does not allow an accurate numerical estimation of the frequency of problems with trainees to be established; this issue is addressed in more detail in study two.

Combining the main results with these methodological issues, I believe that the weighted results of this study are: **that there is evidence to suggest that problems do exist with general practitioner trainees, and that these will often be attitudinal issues; and that the view of trainers on the format of a report form has been established.**

6.1.3 Issues arising

Trainers suggested two main options for recording the judgement of the trainer on the evidence obtained - namely a "yes/no" format or a "yes/no/don't know" format; a rating scale format was rejected. Whilst the latter of these two looks attractive from the point of view of offering the greatest freedom, it is associated with two problems. Firstly, it allows trainers to sit on the fence; in summative assessment this should not be acceptable - trainers should undertake enough assessment, with each criterion in mind, to be able to come to a judgement about whether or not the criterion is being met. Secondly, it would allow trainers to miss out on the assessment of certain sections of the report, simply indicating this by entering a "don't know" response. Bearing in mind that, particularly in the assessment of items dealing with attitudes, a trainer's report would not be corroborated by any other component of summative assessment, this is again unacceptable. It would therefore seem that the most credible answer format to use is a simple "yes/no" format.

This conclusion has two particular consequences for the other studies. Firstly, the use of a yes/no format requires that the criteria developed for the referencing of the assessment are sufficiently specific to allow simple yes/no judgements to be made. Secondly, whilst the use of such a format will prevent the risk of the "centre" effects seen with rating scales the potential risk of the "halo" effect does remain (i.e. that each item in the assessment will not necessarily be judged independently of other items); this emphasises the importance of assessing the discriminatory power of the report form within the field test.

The suggestion that trainers should be able to involve others associated with training to be involved in the assessment is important. Firstly, the involvement of course organisers and other trainers should facilitate the calibration of trainers in their assessments - as some groups emphasised a trainee should only fail the trainer's report after the trainer has discussed their interpretation of the criteria with others who are also involved in the interpretation of the same criteria. Secondly, many trainers recognised the potential organisational difficulties associated with observing trainees, particularly with clinical skills such as intimate examinations; the suggestion of the use of others involved in training (for example consultants or practice nurses) might enable these issues to be handled more sensitively, from the perspectives of both the trainee and the patient. Nevertheless, it would be important for trainers to ensure that the evidence used and the judgements made by others were appropriate for adequate assessment - for example it is no more acceptable for a consultant to say that a trainee performs a vaginal examination adequately purely by discussing the techniques involved with the trainee than it would be if the trainer were to undertake the assessment by discussion alone.

6.1.4 The findings in context

Cronbach outlines two philosophical approaches to the development of tests designed to test attributes in people (Cronbach, 1964) - the psychometric approach and the impressionistic approach. The psychometric approach is the "behavioural corollary of the Thorndike principle of physical science - namely that if a thing exists, it exists in some amount", and that "if it exists in some amount, it can be measured" (Cronbach, 1964). This approach consequently focuses on attempts to obtain numerical estimates of the attributes of the assessee. The impressionistic approach relies on the development of a comprehensive descriptive picture in which "a sensitive observer looks for significant

cues by any available means and integrates them into a total impression; the impressionist is not satisfied with knowing how much of some ability the person has, but asks how the subject expresses his/her ability, the errors he/she makes and why" (Cronbach, 1964). Cronbach (Cronbach, 1964) and Chauncey and Dobbin (Chauncey and Dobbin, 1963) argue that each approach has merit, and each has limitations. Cronbach concludes by stating that "the measurer must fall back upon judgement whenever he applies information from scores .... whilst the portraitist cannot ignore the accuracy of facts that psychometric testing provides" and, as a consequence, he argues that neither approach should be used to the exclusion of the other (Cronbach, 1964).

A trainer's report could follow either one of these philosophies. A report form based on a rating scale adopts a classical psychometric approach; one in which performance is simply described adopts a classical impressionistic approach. The format suggested by trainers in study one sits somewhere between these two extremes - a form of measurement is being made ("yes" or "no"), but, through the use of judgement against a clearly-defined criterion, the approach could be described as descriptive (e.g. yes, my opinion is that the trainee is able to recognise common ear complaints). What are the consequences of such an approach?

In his review Cronbach listed some perceived advantages and disadvantages of the two approaches (Cronbach, 1964). In table 6.1 (overleaf) his conclusions have been taken, and supplemented by my own perceptions, to clarify the likely advantages and disadvantages of the two approaches in each of the areas highlighted in chapter three (p.65) as being of particular importance.

Table 6.1: Perceived strengths and weaknesses of two approaches to a trainer's report (adapted from Cronbach (Cronbach, 1964)).

Property	*Psychometric approach*	*Impressionistic approach*
Content validity	Standardised approach to contents ensures that all content areas likely to be covered following agreed weighting	Reliance on individual may mean that content cover may be variable
Predictive validity	Numerical approach enables comparison with other measures	More difficult to compare with other tests
Overall validity	Depends on attributes that are actually measured (i.e. on the construct validity)	Likely to be high provided portrait is consistent
Stability	Apparent objectivity may improve stability	Apparent subjectivity may reduce stability
Applicability and feasibility	Dependence on measurement likely to increase risks of halo (assessor marks all items same) and central effects (assessor does not use extremes of rating scale)	Reduced risk of central effect. Risk of halo effect persists
Curriculum effects	Measurement may increase "extrinsic reward" effect (p.33). Assessee may attempt to shape curriculum to avoid problem areas	"Extrinsic reward" effect less likely. Assessee may still attempt to shape curriculum to avoid problem areas

In summary I would suggest that a psychometric approach would be favoured on the grounds of curriculum validity, predictive validity and inter-rater reliability, whilst an impressionistic approach would be favoured on the grounds of overall validity, applicability and curriculum effects.

It can be seen from this analysis that the position is finely balanced; an approach that follows one of these two approaches strictly to the exclusion of the other is likely to be disadvantaged; an approach that combines elements of both of these philosophies is likely to be most successful. The approach suggested by the trainers interviewed in study one does provide such a combination.

6.1.5 Conclusions

The results of this study demonstrate that the group interview can be an effective tool for obtaining the views of a large number of individuals. The possibility of the interviewer influencing the outcomes does remain - if this study were to be replicated the use of corroborators (for example audiotapes or additional observers) should be commended in order to enhance the reliability of the conclusions drawn.

The results do provide qualitative evidence that concerns about trainees lie most frequently in the area of attitudes; because of the insensitivity of the other proposed components of summative assessment to this aspect of the practice of the general practitioner trainee this finding supports the need for a trainer's report within this summative assessment process.

The results indicate that trainers would prefer a simple answer format and it is concluded through an analysis of existing work that a simple yes/no format offers a balance between psychometric and impressionistic approaches. An analysis of the advantages and disadvantages of these two approaches suggests that, in this instance, a balanced approach is helpful. Trainers are keen for others involved in training to be involved in the assessment process. They support the use of criterion-referencing and of adequate documentation of the evidence on which the judgement is based. These results provide a basis for the structure of the report form.

6.2 The content of a trainer's report (studies two and three)

6.2.1 Main results

In study two the views of trainers have been used to select the content of a trainer's report. Attributes have been rated both according to their importance for independent general practice and according to which assessment instruments would be most suitable for assessment of that attribute. In study three the degree to which doctors who have recently completed vocational training agree with this choice has been assessed. This study demonstrated that there is broad support for the proposed contents of the trainer's report from doctors who have recently completed vocational training although they do have some concerns about the role of a trainer's report. Study two also demonstrates that a considerable body of trainers have had concerns about trainees sufficient to question whether or not those trainees are fit for independent practice; this occurs at a much higher rate (around 3%) than the current rate of failure (p.31).

6.2.2. Methodological issues

The results of study two do demonstrate that it is possible to select the content of an assessment instrument on the basis of the view of those who will ultimately be involved in the assessment. The method chosen for study two has two particular strengths. Firstly, the size of the sample of trainers used in the study and the high response rate mean that that the results of study two can be considered to be representative of the views of trainers. Although no information is available on the specific reasons for non-response, there is no evidence of bias in favour of a particular age-group or gender amongst respondents. Secondly, by basing the questionnaire on existing work, it is likely that the full range of potential attributes has been considered. Consequently study two does offer a comprehensive, evidence-based blueprint for the contents of a summative process as a whole, and for a trainer's report in particular.

Nevertheless, there are some significant limitations posed by the methodology chosen. Firstly, by basing the questionnaire on published work, it is inevitable that the contents of the trainer's report will, to a very great extent, be constrained to conform to the descriptions of the nature of general practice described in that work. Although the questionnaire did include a section in which respondents could indicate additional attributes for inclusion it is highly likely that their thinking had already been constrained by the questions they had already answered. If this work were to be replicated my view is that an approach which enabled the respondents to think freely would be helpful - for example through the use of initial interviews to define the broad headings under which attributes should be considered, followed by a wide consultation exercise in which contributors were asked to indicate their ideas under each of these headings. Secondly, the sample used in study two means that the content of the trainer's report will be highly

influenced by the views of only one group with a stake in the assessment process. Although study three has provided the opportunity to consider an alternative view, the use of doctors who had recently completed training means that the content of the trainer's report has been determined solely from the viewpoint of two sub-groups of general practitioners. One risk of considering only the views of doctors is that they may provide a skewed perspective of what is important for independent general practice - in particular, the view of the public on the importance of attributes has not been included. Although it may be reasonable to argue that it would be difficult for members of the general public to make judgements about the relative importance of all aspects of the work of a general practitioner when their own experience of general practice might be limited or skewed, the absence of any input from the users of the service must be considered to be a serious limitation. Indeed precedent has been set for the involvement of patients in the selection of contents for an assessment instrument in the context of general practice (Cox and Mulholland, 1993); this process looked solely at the content of general practitioner consultations (a process in which the patient has a particularly close involvement). The choice of two sub-groups of general practitioners as the sampling frames means that the results can not even be considered to be automatically representative of the views of all doctors. This limits the use of these results for any activities other than summative assessment for general practice - for example they should not be directly used for the purposes of determining the content of an assessment process for recertification purposes.

In estimating the frequency with which trainers have had significant problems with trainees, the method used for study two does introduce the risk, common to retrospective studies, of recall bias (Streiner and Norman, 1995) - namely that the

apparent rate will be influenced by the extent to which trainers can recall experiences, with the consequent risk that recent experiences are more likely to be recalled accurately. It is difficult to be certain of the effect of recall bias - the figure may be inflated by trainers who have recently had difficult experiences with trainees which, on reflection, might be considered to represent acceptable performance; conversely, the figure may be reduced by more distant concerns being diluted by the effects of time. Consequently, the estimate that results from study two should be considered no more than indicative of the likely rate of problems.

Study three also has methodological strengths and weaknesses. The response rate of 73% is encouraging, particularly when it is borne in mind that doctors who have recently completed training may be difficult to track down as they will often be mobile (Johnson et al. 1993). Despite this good response rate, the confidence intervals listed in tables 5.3.1 and 5.3.2 are large. This has occurred because, in the sample size calculation, the *ad hoc* estimate of the proportion falling into the desired category was 0.9. Although this estimate might be acceptable for confirmation that the items chosen by one group were also considered important by another group, a number of other results were sought within this study. It would have been preferable to have used 0.5 as the *ad hoc* estimate as this would have provided maximal sample size estimates; the sample size needed would have been approximately three times larger. This emphasises the compromise that is necessary when trying to balance feasibility with minimisation of error in academic studies.

Another major methodological issue that arises in study three concerns the most appropriate way of considering the degree to which another group supported the view of

trainers. Firstly, although there were good reasons for the choice (p.104), because the questionnaire for study three was based on the outcome of study two, it is inevitable that the results will be restricted to the attributes arising from study two and, thereby, will again be restricted to the thinking represented in existing work. Secondly, whilst table 5.3.1 (p.143) indicates a clear separation between those items suggested for inclusion in the trainer's report from the two items judged least important by trainers, the results presented in table 5.3.2 (p.145) are more difficult to interpret; the use of a Likert scale (which allows the respondent to choose to "neither agree nor disagree"), means that the sole use of the proportion of respondents agreeing to the inclusion of the item provides an incomplete picture of the results.

In summary, revising the main results of these two studies on the basis of the methodological limitations of the two studies, I believe that the weighted results of these studies are: **that the contents of a summative assessment process can be based on a sound study of the views of trainers on the importance of previously described attributes for independent general practice; that the views of trainers on the methods for assessing these attributes provide a strong basis for the content of a trainer's report; that there is support for the proposed contents of the trainer's report from doctors who have recently completed vocational training; that no evidence is available from outside of these two sub-groups of general practitioners, in particular from the public; and that there is strong evidence that there are a significant number of trainees about whom trainers have concerns.**

6.2.3 Issues arising

Study two offers two approaches to the selection of content. The first is the selection of contents on the basis of the trainers' ranking of the importance of attributes for independent practice; this offers a basis for the selection of the content of the summative assessment process as a whole. If this method is to be used, some form of cut-off point will be necessary. The effect of choosing various cut-off points is illustrated in table 5.2.2 (p.136). Whilst it is difficult to justify the use of an arbitrary cut-off on any grounds other than the pragmatic balance between feasibility and inclusivity, the suggested cut-off of 70% results in the inclusion of 31 items with all five main subsections being represented by at least one item. This cut-off point would exclude the assessment of many attributes (e.g. the use of some diagnostic equipment, the provision of preventive care, teamwork, practice management, and research). My view is that, in general, these attributes can reasonably be considered as desirable rather than crucial for independent general practice - for example whilst it is highly desirable to work in conjunction with colleagues it may well be possible to provide reasonable care without this attribute - and that a cut-off of 70% does provides an acceptable compromise. Table 5.2.1 (p.134) demonstrates that, whilst trainers believe strongly that the knowledge, skills and attitudes needed for clinical care remain paramount for independent general practice, they also recognise that independent general practice does require more than clinical skills alone - in particular it is important to be able to communicate effectively with patients and to apply ethical principles. Trainers also recognise the importance of general practitioners looking after themselves if they are to survive independently, as exemplified by the need to manage one's own time and to maintain one's health.

The second basis for the selection of content is the views of trainers on appropriate methods of assessment. By providing a view on the contents of the cells of a content/assessment method matrix (i.e. a blueprint), study two offers a basis for the selection of the content of a trainer's report within the summative assessment process. For the 31 items selected on the basis of their importance score the most commonly indicated methods of assessment were the 'written examination', the 'external observation' and the 'trainer's report' with the most frequent response being a combination of one or both of the first two methods in conjunction with the 'trainer's report'. In trying to use these responses as a basis for the choice of items for a trainer's report three approaches are shown in table 5.2.3 (p.137). Whilst the most obvious choice would be to consider the proportion of respondents favouring the trainer's report alone this would fail to take any account of the large number of respondents that indicated a choice of the trainer's report in conjunction with another method of assessment. One alternative is to consider the modal response to the question; this highlights a difference between one item and all other items, but the absence of any numerical data makes it difficult to compare the strength of opinion. The third option is to consider the converse of the first option - namely the proportion of respondents favouring methods of assessment that did not include a trainer's report at all; this option provides quantitative data but also takes into account the possibility of responses in which the trainer's report was only one of a number of chosen options. On balance this method seems to provide the most useful data. Using this approach, of the 31 only one is distinctly different from the others but for four other items more than 25% of respondents felt that a trainer's report should not be part of the assessment. Three of these represent clinical skills; for these items the modal response was 'external observation' in conjunction with the 'trainer's report'. Unfortunately, after the

questionnaires had been distributed, a decision was made that the 'external observation' of practice would be based solely on the analysis of video-taped consultations. This method will not be suitable for many of the practical skills. Because the modal response for these items was the combination of the 'trainer's report' with 'external observation' it would seem reasonable to include them within the trainer's report. The fourth element for which more than 25% of respondents felt that the trainer's report should not be part of the assessment was the element which focuses on NHS obligations and regulations; the modal response for this element was the 'written examination' in conjunction with the 'trainer's report'.

For these reasons it is suggested that the only item of the 31 chosen on the basis of importance that should definitely be excluded from this trainer's report is the assessment of cardio-pulmonary resuscitation although consideration might be given to excluding the item on NHS obligations and regulations from the trainer's report if it is to be consistently covered within the written examination. The results in table 5.2.3 suggest that overlap between the contents of the assessment instruments is likely and may be desirable. This issue is considered in detail in section 6.5.1.

Whilst, in study three, it is reassuring to find that more than 85% of doctors who had recently completed their vocational training agreed that the 31 items proposed for inclusion in the report were attributes that were needed in general practice, it is of concern that when the use of a trainer's report to assess these items was considered the proportion agreeing that this approach was reasonable was rarely greater than 75%, although for 26 of the 31 items ultimately proposed for inclusion significantly more agreed that it was reasonable to assess the item by means of a trainer's report than

disagreed and for none of these 31 was the converse true. For the other five items there was no significant difference between the proportion agreeing and the proportion disagreeing. Three of these five items concerned intimate examinations and one concerned a skill that is only used on occasions by general practitioners (giving an intravenous injection); the fifth item concerned the effect of the trainee's physical and mental health on their ability to work as a general practitioner.

The freetext comments may provide some of the clues to the basis of these findings. The most common comment concerned the effect of the trainer-trainee relationship on the assessment process; many trainees may fear that a poor relationship with their trainer may result in a very unsympathetic view being taken of legitimate absence on sickness grounds. Similarly, a number of respondents commented that some skills, particularly clinical skills, may be more appropriately assessed during the hospital component of vocational training. This may be particularly appropriate for intimate examinations and intravenous injections. It also supports the views of the trainers, highlighted in study one, about the use of others as assessors. In the long term it may become apparent that, under the current vocational training regulations (Anonymous, 1979), just as a structured trainer's report will inform the signature on the VTR1 form (the certificate of satisfactory completion of the training year), similar reports may be needed to inform the completion of the VTR2 forms (the certificates of satisfactory completion of hospital posts for general practitioner training). The freetext comments also highlight general concerns that these doctors had about the place of a trainer's report in summative assessment. When interpreting these concerns it is important to remember that the doctors involved in this study were given no information as to exactly how the report form might be completed by a trainer, and that this study was partly contemporaneous

with study two. Although some of their concerns might be answered by other findings from the research studies (for example the use of others in the assessment, and the development of specific standards) the results do suggest that the experience of recent trainees has rendered them sceptical about whether a report provided by the trainer can adequately consider some aspects of the trainee's skills or attitudes. This highlights the need for a field test (study five) to see if some of their concerns have been adequately addressed. It may also highlight the need for the continuing development of the report to ensure that these concerns are fully addressed. This is considered in detail in section 6.5.3.

The results of study two also show that one-quarter of trainers responding to this study have considered not signing up the VTR1 form. Based on a mean training experience of almost eight years, the finding suggests that about 3% of trainers per year will consider not signing the VTR1 form. Compared with the number of VTR1 forms currently not being signed (Joint Committee on Postgraduate Training for General Practice, 1992; Joint Committee on Postgraduate Training for General Practice, 1993; Joint Committee on Postgraduate Training for General Practice, 1994) the results presented here suggest that for every VTR1 form not signed trainers consider not signing the forms of another 13 trainees. Whilst there may be a number of reasons for this disparity (e.g. trainers being concerned about dooming another doctor's career, pressure from others to sign the doctor up, or that failure to sign up may be seen as an indictment of the training), this result does highlight the need for a revision of the current system, in particular to help trainers in making the difficult decision as to whether or not a trainee is yet ready for independent general practice.

6.2.4 The findings in context

The methodology used in this study offers a novel approach to the setting of the content of an assessment process. For most examinations the content is set by an examination board (e.g. the content of the national assessment tests for children in the U.K. was set by a working group for the Department of Education and Science (D.E.S. 1987), and the examination board of the Royal Australian College of General Practitioners set the content of their membership examination (Fabb and Marshall, 1983)). The use of expert panels to select the contents of an assessment process has one particular disadvantage - that "subject experts will specify different groupings of content to represent their view of the subject" (Willmott, 1978). In this study this risk is minimised by seeking and combining the views of a very large, and representative, group of experts. The results of the content validity study suggest, within the limitations imposed by the methodology (p.174), that the approach used can result in the selection of valid contents.

Many authors place considerable emphasis on the issue of the weighting given to the results of each instrument in an assessment process that combines the outcomes of a number of instruments (Cronbach, 1964; Ward, 1980; Thorndike, 1997). The recurrent appearance of the trainer's report as the selected instrument to assess the attributes judged most important (table 5.2.3, p.137) suggests that considerably greater weighting might be given to the trainer's report than to other instruments. I believe that this would be wrong on two grounds. Firstly, the strength of the trainer's report as an assessment instrument will always be undermined to some extent by the subjectivity of the assessment; all the other instruments involve marking by those not involved directly in the training of that trainee and are consequently objective. Secondly, some intrinsic weighting already exists. Whilst the written test will result in a simple pass or fail, for

both the written submission (Lough et al. 1995) and the consultation analysis (Campbell et al. 1995) a fail will occur if any one of a number of criteria are not met (five criteria for the written test, four for the consultation analysis), these criteria being applied to a single written submission and to at least three consultations. I believe that this results in a crude ratio of weighting of 1 for the written test, 5 for the written submission and at least 12 for the consultation analysis. For the trainer's report failure for any of the 31 items would result in failure of the report form; consequently this component has a crude weighting of 31. This balance is a not unreasonable reflection of the views of trainers on the relative contributions of the four instruments to the whole assessment process. Consequently I do not believe that additional attempts to introduce weighting to the results of the instruments are necessary.

The proportion of trainers who indicate having had significant concerns about their trainee is very similar to results obtained in the one region with experience of undertaking summative assessment (Campbell and Murray, 1996). This suggests that these results are generalisable and that a reliable estimate for the frequency with which trainees are likely to fail the trainer's report component is in the region of 3% per year.

6.2.5 Conclusions

The results of study two provide strong support for a revision of the way in which trainers contribute to the certification process at the entry point to independent general practice. It is likely that the failure rate in this process will be in the region of 3%.

The results of studies two and three demonstrate that it is possible to use the views of a large number of stakeholders in the development of the content of an assessment

instrument and provide an evidence-base for the development of a blueprint for a process of summative assessment in general practice. Whilst some constraints on the content of the process will occur as a result of basing the selection process on attributes already described, such a process is likely to ensure that the blueprint is comprehensive. Within that blueprint, evidence is provided on the relevant place for each of the four assessment instruments proposed.

Although study three does provide some evidence for the validity of the proposed content of a trainer's report, the absence of any input from outside of the profession itself must be considered as a shortcoming in the selection of content for an assessment instrument to be use to select professionals whose prime function is to serve the public; this limitation should be addressed if this work is replicated elsewhere.

Overlap with other assessment instruments does occur; indeed, for the reasons outlined in chapter three (p.63) overlap may well be desirable, particularly for some attributes. At this stage of development I do not believe that this overlap implies that substantial revision is needed to the proposed content of this trainer's report, nor do I believe that the introduction of an arbitrary weighting to the results of the four proposed instruments is necessary. This is considered further in section 6.5.1.

Doctors who have recently completely training do have some concerns about aspects of the trainer's report. Whilst the importance of some of these concerns is likely to become apparent from the results of the field test, the continuing development of the trainer's report should aim to answer further these concerns where necessary.

6.3 Setting standards (study four)

6.3.1 Main results

Study four demonstrates that a formal consensus exercise can be used to set standards; the addition of a consultation exercise resulted in only minor alterations to the standards set in the consensus exercise. The outcome of these two exercises is a small number of absolute standards for each of the content areas proposed. These standards are consensus minimum standards. Because of the inclusion of terms such as "repeatedly" and "appropriately" within the wording of many of the standards, it is apparent that trainers will need to use their professional judgement in the interpretation of the meanings of these terms. Evidence for all bar one of the standards can be obtained through direct observation of performance whilst for all proposed standards it is possible for the assessment to be made by someone other than the trainer.

6.3.2 Methodological issues

The method chosen for setting standards for the trainer's report was designed to meet the requirement for the standard-setting process to be fair and unbiased (Bowmer, 1994; Dauphinee, 1994). The presence of delegates at the consensus conference from more than three-quarters of the areas served by a Director of Postgraduate General Practice Education (21 out of a possible 27) does substantially reduce the risk of a particular regional bias to the standards set. Because all of the trainer delegates were experienced trainers, and because all Directors of Postgraduate General Practice Education were asked to be involved in the consultation phase, the requirement for the standards to be set by "an adequate number of judges who are knowledgeable (some of whom are experts or leaders in the field)" (Bowmer, 1994) has also been addressed.

At the consensus conference the need for the standards to be absolute rather than relative, and minimum rather than optimum, was reinforced. Examination of the resulting standards (appendix 5.2) demonstrates that the standards set are absolute rather than relative. Similarly observation of the group processes demonstrated that the groups did focus on minimum standards and avoided trying to set optimum or gold standards.

There are two main weaknesses to the method used in this study. The first is the limited perspective gained in the setting of standards. There has again been no input from outside of the profession in the setting of these standards. Whilst this strictly adheres to the current principles of professional self-regulation, the arguments presented on p.32 suggest that professional regulation must become much more sensitive to the views of the public served by the profession. If this work were to be replicated this weakness should be addressed. The second weakness of the method is the use of only one person to interpret the standards written on the worksheets. This introduces a risk of observer bias in the interpretation of standards. Whilst this is somewhat diminished through the use of a consultation phase, the risk could be further reduced through the use of two or more interpreters with their interpretations being examined for inter-observer consistency.

In summary, I believe that the weighted results of this study are: **that consensus minimum standards have been set for all attributes proposed for assessment in this trainer's report, but that they are limited to a view from within the profession; that assessments against all standards can be made by someone other than the trainer but that trainers will need to use their professional judgement in the interpretation of many of these standards.**

6.3.3 Issues arising

Having developed standards for each of the items proposed for assessment by this trainer's report, three particular issues relating to the standards warrant further consideration.

The first issue is whether or not the standards have truly been set at the minimum level for entry to independent general practice. Although the absence of major modifications during the consultation phase does provide some support for the view that they do represent appropriate standards, it must remain a matter of opinion as to whether the standards set are truly the minimum acceptable standards for entry to the profession. Further work to corroborate or refute the fitness of these standards would be worthwhile.

The second issue is that, if the standards set for the assessment are truly minimum, there may be a risk that the whole curriculum will become minimalist. It was argued in chapter three (p.54) that there is a clear difference between having a minimum standard and having a minimalist approach. All examinations will have a minimum standard - that is, a standard which denotes the boundary between success and failure; what is of concern is when the focus on minimum standards drives both trainees and trainers to aim no higher than the minimum standard i.e. the approach becomes minimalist. My view is that minimum standards do not necessarily predicate a minimalist approach. On p.55 it was argued that an assessment process using minimum standards can be associated with an optimal approach. At an individual level, if trainers were to focus their formative assessment on the content areas contained in the trainers report, they and their trainees

should be able to recognise early on in training those areas on which effort needs to be concentrated to ensure that the trainee will pass summative assessment; once they are confident that this assessment can be passed then it should be possible for the majority of the training year to focus on higher levels of performance in these areas - once the trainee is known to be safely on the "pass" side of the boundary posed by the minimum standard, there is no reason why the focus of the training should not become considerably higher.

The third issue arises from the need for trainers to exercise professional judgement in their interpretation of the standards. When using the term "repeatedly" or "persistently" how many observed errors should cause failure of an individual item and thereby of the trainer's report? What is of most concern is unsatisfactory performance that is likely to continue once a trainee enters independent practice; this is more likely to happen if it has been seen to happen repeatedly during the training year (Asher and Sciarrino, 1974). Conversely trainees should not be failed on the basis of a single occurrence; failure without a chance to improve performance as a result of learning from mistakes seems to contradict natural justice. I believe that it is unwise to prescribe an exact number of errors that should result in failure - what is needed is for a judgement to be made that takes into account such factors as the seriousness of the error, whether the error continues after advice, and the balance between the frequency of error and of success in that particular aspect of performance. Consequently while these terms may look imprecise I believe that it is important that trainers use their judgement in interpreting these standards. The consequence of the reliance on judgement is that the chances of erroneous judgements being made must be minimised. This is considered in more detail in section 6.5.4.

One particular issue concerning methods for standard-setting arises from this study. The small number of changes that resulted from the consultation exercise must call into question the value of adding a consultation phase to the consensus conference. The inclusion of a consultation exercise does seem to offer four significant advantages: it provides a validation exercise in which standards can be reviewed within a structured framework before being released more widely (the absence of major changes in this study could be interpreted as indicating that the standards did have face validity); in a more political sense it preserves credibility in that it offers the opportunity for major influencers to be involved in the process without the consensus conference being seen to be dominated by those whose day-to-day experience of trainees may be limited; similarly, by including educational leaders, it offers the opportunity to increase the perception of ownership of the standards by those leaders; finally, by including those who attended the consensus conference within the consultation process, it allows attenders the opportunity to revise their input after reflection (thereby further limiting the influence any one individual can have on the final outcome). If the work of this study were to be repeated my view is that these are significant advantages in ensuring the professional credibility of the standards set.

6.3.4 The findings in context

Within the work published about assessment considerable confusion appears to exist about exactly what the term "criterion-referencing" means. Work that refers to criterion-referenced tests often uses criterion as a term to define the exact domains against which the assessee will be assessed - criterion-referencing is concerned with anchoring the assessment to highly specified content areas (Popham and Husek, 1969; Thorndike,

1997). The use of the term in this way perhaps explains why such authors place such importance on content validity for criterion-referenced tests and is perhaps best replaced by the term "content-specific testing". Elsewhere the term is used in relation to the standards used for the assessment (Glaser, 1963; Glaser, 1971) - criterion-referencing is concerned with measurement against "absolute standards of quality" (Glaser, 1963). The standards take the form of "classes of behaviour that define different achievement levels (which) are specified as clearly as possible before the test is constructed" (Nitko, 1971); this approach is perhaps best described by the term "criterion-referenced measurement" (Glaser, 1963). Although this trainer's report is a content-specific test that uses criterion-referenced measurement, many instruments will not fulfill both of these definitions. I believe that clarity is required in the assessment literature in the use of the term "criterion-referencing" and that the alternatives of "content-specific testing" and "criterion-referenced measurement" do provide greater clarity.

Support for the use of judgement in the interpretation of standards is offered by Tonesk who argues that because "there is a danger that as evaluation is made more objective, it also becomes less meaningful", "there is a place for ... judgement throughout the ... process" (Tonesk, 1983), and Wolf who argues that "assessors do not simply 'match' candidates' behaviour to assessment instructions in a mechanistic fashion" (Wolf, 1995). Support for the choice of wording of the standards is also provided by Ward who argues that the phrasing of the items should reflect behaviours that are directly observable (Ward, 1980).

Because the aim of this study was specifically to set minimum standards for entry to independent general practice it is difficult to compare the outcomes with the standards

described in similar supervisor reports (Rakowski, 1990; Preece, 1993). However, it is clear that, in contrast to these other two trainer's reports, the standards used in this report have been set in an overt way.

Whilst criticism can be leveled at the use of a consensus conference and a consultation exercise for the purposes of standard-setting as an alternative to the setting of standards entirely based on published evidence (Farmer, 1991), the paucity of evidence about appropriate minimum standards for independent general practice does mean that a scientifically pure evidence-based approach would be currently impossible. Some support for the approach used in this study is provided by the similar consensus approaches taken to standard-setting for two of the other instruments proposed for the summative assessment process for general practice - namely the assessment of consultation skills using video-taped recordings (Campbell et al. 1995) and the assessment of written communication skills using an audit project (Lough et al. 1995).

The approach outlined in this study does offer a starting point for the setting of standards against which further attempts at standard-setting can be compared.

6.3.5 Conclusions

This study has resulted in the development of consensus standards for use in a trainer's report. These are absolute standards directed at the minimum level of performance for entry into general medical practice, although it has to be accepted that they are formed on the basis of a limited perspective. Their application, either by trainers or by others involved in training, will require judgement. Consequently some form of quality

assurance programme should be considered obligatory if public confidence in this approach to professional self-regulation is to be maintained.

If this approach to standard-setting was to be replicated I would advocate the inclusion of a consultation phase even though, at first sight, it may not appear to result in significant revisions to the proposed standards.

Clarity in the use of the term "criterion-referencing" in the assessment literature is needed. I would advocate the use of the terms "content-specific testing" and "criterion-referenced measurement" for the two uses currently made of this term.

6.4 Field testing the report form (study five)

6.4.1 Main results

The principal results of study five are that the report form does have discriminatory power, is feasible, and is reliable with high levels of inter-rater agreement being found for all items in the trainer's report (without evidence of systematic bias or collusion between trainers), and a kappa coefficient indicating agreement of moderate strength on the overall outcome of the report. Certain items (particularly intimate examinations) do cause some problems.

6.4.2 Methodological issues

Three major methodological difficulties arise with this study: difficulties resulting from the use of small samples; concern about how representative of all trainers the sample of trainers is; and difficulties associated with using the Cohen's kappa coefficient to describe the reliability of this instrument.

This study was set up to allow the testing of this trainer's report within the context for which it was designed. This required not only that the report was used over a period of time but also, because of the need to assess inter-rater reliability, the cooperation and time commitment of two trainers as well as the trainee within a practice. It is therefore not surprising that only 69 of the 511 practices contacted felt able to become involved, particularly when it is borne in mind that at any one time only half of these trainers are likely to have a trainee working with them (Government Statistical Service, 1996) and that, of these, up to one-quarter might be expected to have joined or left the practice during the study period of three months. The greatest number of practices that could have been recruited is therefore likely to be in the region of 180, of whom just under 40% were recruited. Unfortunately, a sample size estimate for the reliability component of the study indicates a need for approximately 250 practices (based on a predicted *kappa* coefficient of 0.5 and a desirable margin of error of 0.1) (Streiner and Norman, 1995). The situation is made worse by the finding of Suissa who demonstrated that when a dichotomous marking schedule is used (in this case pass/fail) the sample size needed to show an effect is increased by a minimum of 50% (Suissa, 1991), resulting in a desirable sample size of 375. The desirable sample sizes for the discriminatory power and feasibility estimates would be 384 (based on a desired accuracy of the estimate of +/- 5%) (Mant and Yudkin, 1993). The principal consequence of having to work with a small sample size is that the results of this study can be taken only as a guide of the likely reliability, discriminatory power and feasibility, and that the large margins of error have to be considered when the results are being interpreted and used.

Furthermore, of the original 69 practices willing to be involved, only 56 provided at least one report form. The information available from 8 of 13 practices who did not complete the study indicated that the reasons lay in difficulties within the practice rather than with the report form itself. Nevertheless, it is possible that for the other five practices the report form itself was the cause of their non-completion. The absence of data from 7-8% of participating practices, whilst not being likely to render the conclusions reached from the remaining practices completely invalid, does further lessen the strength of the conclusions reached.

The second concern with this sample lies in the possibility that the trainers involved are not representative of all trainers. Unfortunately it is not possible to give any accurate indication about the extent to which these trainers are representative of their peers as little information is known about trainers as a whole.

Finally, although the *kappa* coefficient is widely used as a numerical indicator of reliability it does have considerable limitations. The principal difficulty with the use of the *kappa* coefficient arises from its dependence on the prevalence of the attribute being measured (Brennan and Silman, 1992); the exact value becomes less meaningful when one value (in this case a "pass") is much more likely than the other value. Furthermore, there is some evidence that suggests that the use of dichotomous scales further reduces the value of the *kappa* coefficient obtained; in a study of certification examinations a coefficient of inter-rater reliability of 0.76 obtained when the full range of original scores was used fell to 0.69 when the scores were converted to pass-fail decisions (Streiner and Norman, 1995). These reservations need to be borne in mind when interpreting the value of *kappa* presented.

Based on these methodological limitations, I believe that the weighted conclusions from study five are **that the results suggest that the report form has acceptable levels of inter-rater reliability and feasibility, and is able to discriminate a group of doctors. However, because there are substantial weaknesses that result, in particular, from the sample size of the study that tested these properties, the collection of further data would be justified.**

6.4.3 Issues arising

Reliability

The results demonstrate high levels of inter-rater agreement, but only moderate agreement when the *kappa* coefficient is used as the measure of agreement. Thorndike argues that "for a test that is being used for the single go-no go decision .... the percentage of consistent decisions seems to be a reasonable index of reliability" (Thorndike, 1997). The rationale for this argument is that reliability indices that are based on just four comparator cells (i.e. a two-by-two table of pass vs. fail) are vulnerable (because of the mathematical manipulation required) when cells contain numbers of different orders of magnitude. He goes on to argue that "if we must make some decision ... we will do so in terms of the best information we have - however unreliable it may be - provided only that the reliability is better than zero, in which case we have no information" (Thorndike, 1997). I would draw two conclusions from his rationale: firstly, it is reasonable to draw some conclusions solely from the demonstrated levels of agreement between trainers; secondly, the information available from this study is that the reliability is better than zero.

Overall validity

It must be emphasised that the rate of failure in this study should not be taken as an estimate of the 'absolute' failure rate that might be expected if the system were fully operational - the observed rate would be expected to be high because the design of the study deliberately included doctors who were only part way through their general practice year. Similarly, although the results provide some indication as to which items are most likely to result in failure, they only apply to trainees who are part way through their training period and should not be directly generalised to the likelihood of failure for trainees reaching the end of their training.

The finding that failure occurred in 19 of the 31 items provides no evidence, at this stage, to support a radical reduction in the number of items; further evidence on the relative discriminatory power of items for trainees completing training should be sought before considering such action. Specific review of the items causing failure demonstrated that three trainees would have been failed by one trainer only. There may be a number of reasons for this - different trainers may have had different experiences of the trainee; different trainers who have had the same experience of the trainees may interpret standards differently, with "hawks" unfairly failing trainees, and "doves" not failing trainees whose performance is below minimum standards. This finding adds further emphasis to the need for an adequate system of quality assurance for a trainer's report. Because of the possibility of false positives occurring (i.e. doctors being failed who should not have been), it is essential that any trainee who is failed should automatically have their performance reviewed. In particular this may help to clarify not only whether or not a problem exists, but also, where it does exist, whether the problem is localised or global. Conversely it is also crucial that the possibility of false negatives (i.e. doctors

passing who should not have done so) is considered to ensure that public confidence in the system can be maintained. These issues are explored in more detail in section 6.5.4.

Feasibility

The results demonstrate that the proposed report form is feasible, with less than 6% of lead trainers unable to assess any one item. When interpreting these results it must be remembered that for four of the 56 practices one or other trainer did not return a completed form; although it can not be accepted as certainty this may indicate difficulties with the feasibility of completing the form. Nevertheless the low level of inability to assess items is particularly encouraging when it is remembered that the study took place over only three months. Whilst the use of only the forms returned by "lead" trainers as the denominator might be construed as attempting to increase the level of feasibility (as these trainers might be expected both to have greater experience of the trainee and to have a greater investment in the assessment of the trainee), this measure was used because it reflects how the report form is actually designed to be used (i.e. completion by a single trainer).

The main difficulties lay in the assessment of intimate examinations. As many trainers suggested, in the long-term it may be that these assessments would be better undertaken during the hospital component of vocational training or by the use of mannequins. The availability of standards that allow assessment by colleagues of the trainer (study four) should enable such a sharing of assessment to occur.

Other issues

It is encouraging that within the spontaneous freetext comments an equal number of respondents commented positively on the report form as commented negatively, with an additional number recognising the strength of a report such as this when a problematic trainee was being assessed.

Respondents noted problems with the guidance notes, and a large number of suggestions for improvement in the report form design were made. Two comments in particular need to be addressed. Firstly, there was concern that some items appeared repetitive, resulting in a report that was time-consuming to complete. In the long-term this might be most appropriately addressed by considering the discriminatory powers of each item when the report form is used by large numbers of trainers on trainees nearing the completion of their training - it may become clear that some items cause failure so rarely that their inclusion adds nothing to the overall report. The frequency of this comment also suggests that, if desired, a test of the internal consistency of this report form might be possible. Secondly, a number of trainers raised the possibility of having a two-tier system of assessment, with this report form only being used when concerns had already been raised. There are a number of objections to this approach. Firstly, such a system has not been tested; it would be similar to what is currently available, and concerns have already been expressed about whether that system does actually work (study two, p.180). Secondly, any initial screening process would have to be rigorous, and may consequently have to be quite detailed and not substantially different from using this trainer's report on all trainees. Thirdly, for certification purposes it is probably fairest that all trainees are submitted to the same test to reduce the stigma associated with being submitted to a second more detailed assessment.

6.4.4 The findings in context

There is a very strong body of opinion that new assessment instruments should have demonstrable levels of validity and reliability (Cronbach, 1964; Hudson, 1973; Ebel, 1979; Ward, 1980; Thorndike, 1997). Nevertheless at least two authors counsel caution in the pursuit of these properties. Raven cautions against being wooed into a false sense of security through the apparent attraction of numerical measurements of the quality of a test (Raven, 1991). He argues that the requirement to select only those instruments with demonstrably high levels of validity and reliability may result in only instruments that lend themselves to such measurement being designed; this may detract from other, less quantifiable, qualities. Thorndike argues that traditional measures of reliability do not work well for criterion-referenced measurements because there is "little variability in the set of scores (which) .... tends to yield low reliability coefficients" (Thorndike, 1997). On balance, I believe that an assessment instrument that is to be used for a high stakes test must have levels of reliability that have been demonstrated using widely accepted tests, even if this does constrain (at least to some extent) the design of the instrument. The evidence of study five is that, with reservations, the trainer's report does have demonstrable, and probably acceptable, level of inter-rater reliability.

Although the testing of the Exeter Teaching Practice Schedule was strictly a concurrent validity study (because the two groups were not both using the report form) it is interesting that all seven students identified as performing poorly by the supervisors using the report form were also identified by the teaching staff in the schools in which they were placed; similarly all those who passed on the supervisor's report were also given a pass grading by teachers (Preece, 1993). These results, using a similar report form but in

a different educational setting, confirm that high levels of inter-rater reliability are potentially achievable with reports provided by a trainer.

It has already been argued that an important measure of validity is overall validity (p.61). Further support for the use of overall validity as a crucial measure of validity is provided by Thorndike who argues that "what is validated is not the test itself but the interpretations of the test scores for particular purposes or uses" (Thorndike, 1997). This view is well supported by others (Bean, 1953; Ebel, 1979). For this instrument overall validity has been demonstrated.

On p.194 it was suggested that further evidence once the report form was in wider use would be helpful. Towards the end of 1998 results from a year of use of the final version of this trainer's report were made available (Attwood, 1998). This demonstrated that of 1467 candidates for summative assessment, 42 failed summative assessment (2.9%); of these 42, five failed on the trainer's report alone and 26 failed the trainer's report and one of the other three components of summative assessment (i.e. 31 failures on the trainer's report (2.1%)). These results support the overall validity and feasibility of the trainer's report when used in the setting for which it was intended.

6.4.5 Conclusions

Although there are considerable limitations to the weight that can be put on the results of this study as a direct result of the method used for the study, there is evidence, based on accepted measures, that indicates that the trainer's report is likely to be valid, reliable and feasible in use. This finding has subsequently been further supported by the results of the use of the final version of this trainer's report nationally.

The evidence from the field test does not suggest the need for major revisions to the instrument itself, although revisions to the guidance notes used in the field test are required.

These findings suggest that this instrument offers a suitable first trainer's report for use in a summative assessment process for the purposes of certification at the completion of training for general medical practice in the U.K. The findings of this study also suggest, very strongly, that a quality assurance programme is crucial.

6.5 Overall conclusions and the continuing agenda

Although the range of technical properties sought and tested in the research studies which form the basis of this thesis was limited (for the reasons given in chapter three (p.76)) the research programme presented in this thesis has resulted in the development of a trainer's report form for use as part of a summative assessment process for general practitioner trainees that has demonstrable, and acceptable, levels of those components of validity, reliability and feasibility judged important. The overall aim of the research component of this thesis (p.76-7) has therefore been fulfilled. The resulting final version of this trainer's report is presented as appendix 6.1.

Although there are some aspects of the instrument that are less than ideal (in particular the absence of involvement of members of the public in the development of an instrument designed to assess public servants) this report form does have some features as an assessment instrument which, in combination, seem to be unique. These are: a transparent selection of content based on evidence on both the importance of attributes

for the future vocation and assessibility of those attributes by the particular instrument; minimum standards for criterion-referenced measurement set by means of an overt consensus process; and demonstrable and acceptable levels of overall validity, content validity, inter-rater reliability and feasibility.

The final section of this chapter considers the continuing agenda that arises from this work. An analysis is made of four issues: what would be required to enable widespread implementation of a report form such as this?; what further research is worthy of consideration?; how should the current form continue to be developed?; and how can the issue of a quality assurance programme be best addressed?

6.5.1 Implementation of the report

If a new assessment instrument is to be adopted as part of a national certification test a number of issues need to be addressed.

Firstly it is essential that the assessment instrument is accepted by the professional body with the mandate to control certification - for this report the JCPTGP.

Secondly, the relationship between the instrument and other components of the assessment process needs to be addressed. Because each of the instruments proposed for this process (Joint Committee on Postgraduate Training for General Practice, 1994) was developed independently (Campbell et al. 1993; Campbell et al. 1995; Lough et al. 1995; Campbell and Murray, 1996), there is a considerable risk that the content of the overall assessment will not follow rigorously the assessment blueprint developed in study two. Criteria for dealing with overlap were suggested on p.64. Overlap with the written

test is likely to occur for those elements concerning knowledge about the use of drugs and understanding the obligations of a general practitioner according to the NHS contract and regulations. Whilst I believe that the latter of these can be reasonably assessed by means of a written test, I believe that the very high levels of importance ascribed to knowledge about prescribing by trainers in study two means that a longer-term assessment of performance in this area is justified. Overlap with the assessment of consultation and communication skills is likely to occur in the areas of problem definition, problem management and the use of resources. Because it is proposed that the analysis will only provide evidence on consultation skills from a maximum of two hours of consulting (which is likely to equate to a maximum of 12 consultations) (Joint Committee on Postgraduate Training for General Practice, 1994), I believe that it is justifiable to retain equivalent sections in the trainer's report as assessment in these sections could be based on evidence obtained from a considerably greater number of encounters with patients and colleagues. Finally, overlap with a written submission of practical work is most likely to occur for those items dealing with the trainee's awareness of his/her own limitations and the ability to identify strengths and weaknesses in his/her own performance. Unfortunately, it is quite possible that a written submission would provide no evidence in this area - whilst providing considerable evidence both about the performance of the setting in which the trainee works and about the trainee's numeracy and literacy skills, such a project may give little insight into whether or not the trainee has the aptitude of self-awareness. As a result, it seems reasonable to maintain these items within the trainer's report.

So what effect does the final choice of methods for the other components in the proposals for summative assessment for general practice (Joint Committee on

Postgraduate Training for General Practice, 1994) have on this trainer's report? On balance it would seem that, with the possible exception of the two areas of overlap with the written test, at this stage the proposed trainer's report should remain in its current form. Once experience has been gained it may be possible to remove some items that are being assessed by other methods, particularly if the alternative methods are demonstrably more valid and reliable. If experience over the next few years does demonstrate that this is the case, then it may well be appropriate for the trainer's report to be revised.

The third major issue for implementation is that of ensuring that trainers have the skills to use the form. In the field test it was intended that no input would be made to the trainers involved in the study that would not be made for all trainers when it was implemented; the demonstration that the report is feasible to complete and valid and reliable in use (within the methodological limitations already mentioned) suggests that it should be possible to implement this trainer's report without a massive educational programme. However, I do believe that trainers may need support in two areas. They may need support in making judgements using the evidence they collect; structures will need to be in place that enable trainers to learn from each other (and from experts) about honing their skills in this area. Many trainers may also need support to combine the training and assessing roles; although this might be regarded as a new dilemma to face trainers the evidence from study two (in which many trainers admitted to having considerable concerns about the performance of some trainees when completing the VTR1 form) suggests that this dilemma is not new but has simply become more explicit.

In summary, I believe that, if this report were to be adopted widely as one instrument within a summative assessment process, there is a need for supportive structures to be in

place that enable trainers to reflect and build on their experiences; specific training interventions are not necessary. This means that the development of trainers to support the implementation of this report is likely to be manageable, should not cause particular confrontation with trainers, and is probably best done through local networks rather than through a major national training initiative.

6.5.2 The remaining research agenda

In chapter three potential research questions were selected on the grounds of importance. Of these, two (predictive validity and curriculum effects) were excluded on the grounds that they were not currently feasible. This section revisits these issues, and also considers what other research issues warrant further consideration.

Predictive validity testing

The possibility of undertaking a formal predictive validity study was dismissed in chapter three firstly because, at the time the validity tests were being designed, there was no performance-based re-certification process against which the results of summative assessment could be subsequently compared; and secondly because such a study was not feasible within the time-scale of this thesis (p.73). A predictive validity study would also be difficult if a purist view of a predictive validity study is used in which there is no further influence on the performance of a doctor after the end of summative assessment that is restricted only to those trainees who had failed - a purist approach of such a study would not sit easily with a plan to offer further educational intervention to those who fail a summative assessment process.

It is now worth re-examining the possibility of undertaking a predictive validity study principally because the issue of re-certification at the level of minimum standards of performance is being addressed by the General Medical Council (Brearley, 1996). The rate of entry into the General Medical Council Performance Review Procedures during the first five years after the passing of summative assessment could be taken as a crude measure of the false negative rate of summative assessment, and the numbers could be added to the numbers who have failed summative assessment despite appeal and/or further training (the true positives of summative assessment) to provide a total number of positives. Furthermore, if a more liberal interpretation of a predictive validity study is taken in which the process of appeal and re-training is included as part of the overall process of summative assessment, the true negatives would consist of those who pass summative assessment and who are not entered into the Performance Review Procedures; the false positives would be those who initially fail summative assessment but who pass as a result of either the appeals procedure or re-training but who are not subsequently entered into the Performance Review Procedures. This approach would provide a review of all who had undertaken summative assessment.

Although this approach looks attractive there are a number of difficulties that would need to be overcome if such a study were to be undertaken. Firstly, such a study would consider the combined result of all components of the summative assessment process rather than a single instrument. Secondly, the overall outcome measure being used (i.e. entry into the Performance Review Procedures) is as yet untested; these procedures have yet to be validated as a measure of performance - not only is it as yet unclear that they will measure performance in a valid way, but there must remain considerable doubt about whether all doctors whose performance is poor will actually be subjected to these

procedures. Thirdly, questions might be raised about the acceptability of such a study, particularly around which body would undertake the study and what method would be used; it would be most likely to be acceptable if a national body with a major interest in standards (such as the JCPTGP or the General Medical Council) were to undertake the study, using a methodology that provided not only numerical data but also qualitative data, perhaps based on the methods employed in the critical incident technique (as originally described in relation to the failure of pilots during training (Flanagan, 1954)). On balance, my view is that such a study may become feasible once further information on the validity of the Performance Review Procedures is known, but that such a study would be a major undertaking.

Curriculum effects testing

A study examining the effects of a summative assessment process on the curriculum was excluded in chapter three (p.75) from the proposed research studies on the basis that this effect could only be measured after the introduction of the process on a national basis. The adoption of summative assessment nationally (Anonymous, 1997) provides an opportunity for such a study. A study could include an analysis of the effect of assessment on the relationship as well as an analysis of the curriculum effects of including assessment of this type.

Other questions

A number of other research questions would arise with the introduction of a summative assessment process, which includes a trainer's report, on a national scale; in particular:

- does this trainer's report work when it is applied universally? Initial evidence suggests that it does (Attwood, 1998) but this evidence provides no information about issues that arise from its use.
- is the system cost-effective? This thesis contains no analysis of the costs either of the total summative assessment process in this context, nor of a trainer's report in particular. Any analysis of costs should not only examine such issues as time and financial resources, but should also examine such issues as trainer- and trainee-motivation and opportunity costs (particularly any diminution in the time spent on training).
- are there systems that are more efficacious? For example, would a staged system in which there is an initial screening phase followed by an in-depth analysis of some doctors be as effective but at lower cost?
- to what extent does internal concurrent validity exist between instruments within the summative assessment process? This might offer an insight as to whether the instruments do measure similar attributes and, if so, whether the instruments could be modified to simplify them whilst maintaining comprehensive coverage of the important content of general practice. Ultimately, consideration would need to be given as to whether all components were needed.
- how should the content of the trainer's report be modified to ensure that it is in tune with developments within the discipline? The content basis of this report does, to a major extent, reflect a view of general practice that is up to fifteen years old; because the discipline is likely to have shifted in that time consideration must be given to the way in which the contents of the trainer's report will shift to reflect these changes.
- how should the standards be reviewed to ensure that they continue to reflect the minimum standard acceptable for independent practice? The standards set for this

report reflect the view of experienced members of the profession at one particular point in time. Consideration should be given as to how the standards should continue to be developed to reflect the changing requirements of the discipline. Any such process should also enable an assessment to be made as to whether or not the standards set do truly reflect the minimum standards.

Whilst there may be merit in research programmes in all of these areas, the last two issues focus on the way in which the report form can continue to match a developing discipline; they are essentially developmental issues rather than purely research issues. They are considered in more detail in the next section.

6.5.3 Continuing development of the report

General practice is a profession that is always undergoing subtle change. Indeed, since this project was initiated there has been at least one new attempt to detail the work of the general practitioner (The Nature of General Medical Practice Working Party, 1996). Consequently any instrument that purports to assess the readiness of an individual to work independently within general practice will need to change too.

The contents of this trainer's report could be updated in at least two ways. The first is to undertake a broad consultation exercise, such as the questionnaire survey described in study two, from time to time. This has the major advantage of ensuring that additions (and subtractions) from the content of the report are widely supported by those involved in its use. However it is a resource-intensive exercise and could only be undertaken relatively infrequently; although this might ensure that changes were not made solely as a result of short-term fashion it would mean that the report form would always be some

years behind changes in practice. The second option is for a small group of those heavily involved in general practice education to take responsibility for its continuing development. Whilst this approach would ensure that responsibility for updating was maintained and that changes could be adopted quickly, it does suffer from the major limitation that such changes might be seen by those not involved as merely reflecting the views of an interest group; the objectivity which the first approach offers would be replaced by a strong flavour of subjectivity. Perhaps the most suitable option would be for a small group to take on responsibility for ensuring that continued development did take place, but that any changes in content only resulted from a broad consultation exercise. If either of these options were to be used, two lessons have arisen from the work of this thesis which should be incorporated into such initiatives. Firstly, it is crucial that any consultation exercise also takes account of the views of others involved in general practice including non-training doctors, other primary-health care team workers and patients. Secondly, the exercise should not be confined solely to existing descriptions of the work of the general practitioner; methods would be needed that neither confine respondents nor undermine the comprehensive nature of the assessment that is required in the trainer's report. One approach that might be used to overcome these limitations would be to use group interviews to develop the contents of a questionnaire that is comprehensive but not constrained by existing descriptions which is then used as the basis of a study which again allows rating on the basis of importance and assessibility; both phases should include a wide breadth of opinion.

As the nature of general practice changes not only should the contents of the report be considered for change but so too should the standards. Indeed, it is likely that as the discipline continues to develop both those within the profession and those outside will

demand that standards continue to rise (p.22). Although this could be done simply by repeating the exercise of study four a number of modifications to that process are worthy of consideration. Firstly, if accountability is to be improved, there should be public involvement in the standard-setting. If this were to be done, I believe that considerable difficulties might arise if there were to be mixed professional/lay groups working in the consensus phase; consequently public opinion might be best included using a structured consultation exercise of the type used in study four. Secondly, there should be representation of the General Medical Council in the consensus phase The rationale for this is to ensure that the minimum standards set could be applied to doctors across the whole range of experience - whilst it may be argued that the minimum standards for experienced doctors may legitimately be set at a higher standard, it does not seem reasonable that the standards for experienced doctors should be any lower than those set for novice doctors. Thirdly, those involved in the consensus phase should probably change each time. This would prevent the standards being consistently influenced in a particular direction by particular individuals, and may also reduce the risk of group members feeling obliged to set different standards simply to ensure that new standards exist (an "anti-halo" effect). The timing of the revision of standards may be influenced by many factors, particularly political ones. For simplicity it would probably be most appropriate for the revisions to be made after revisions to the content have been considered.

Although the need for specific weighting of the trainer's report in relation to the other components of summative assessment has been dismissed (p.181), the issue of internal weighting (that is, the relative weighting of items within the report) is worthy of further consideration. Such a mechanism is being considered in the context of clinical skills for

specialist physicians (Dacre, 1996). This involves the division of the skill into individual stages; for each stage a "mark" is derived by experienced clinicians which takes into account both the degree of difficulty that the stage involves and the degree of importance in completing the stage correctly (thus stages of high importance and low difficulty would have to be achieved during the assessment, whilst a proportion of less crucial stages of greater difficulty could be done less well whilst still allowing the doctor to achieve an overall pass). There are a number of limitations to this approach. Firstly, its application may be limited - whilst it lends itself well to the specific clinical skills component of the trainer's report, it would not be easy to adopt this method for most of the other items. Secondly, simple weighting mechanisms may oversimplify the complex makeup of some of the attributes tested in the trainer's report. Thirdly, it requires a 'mark' to be ascribed to the performance; not only was such a method considered undesirable for the trainer's report by those trainers involved in study one, but the ascribing of marks would considerably complicate the completion of the trainer's report and would require considerable training of all trainers to ensure consistent application. Consequently, I believe that the introduction of internal weighting is not currently justified.

6.5.4 Quality assurance

Perhaps the most pressing development issue is the need for a quality assurance programme. If this trainer's report were to be widely adopted it would be crucial that doctors who pass the report are only those who should do so, and that those who fail are only those who should have failed. The first is an issue of public confidence; the second is an issue of natural justice for the doctors concerned.

One approach, analogous to the "Total Quality Management" approach advocated in industry (Berwick, 1992a; Berwick, 1992b), involves a systematic and prospective approach to the minimisation of error. In this context, for example, trainers could be encouraged to observe a trainee on a number of occasions in order to make their assessment on each item of the report, and could be advised to discuss their interpretation of the standards to ensure that their interpretation would be supported by others with experience in training.

Although this approach may help to minimise the chances of an erroneous judgement being made consideration must also be given to the development of some form of quality assurance mechanism that examines whether or not correct judgements have been made. For public reassurance the crucial issue is 'sensitivity' (i.e. that there should be a minimum number of doctors who pass when they should have failed (the false negatives of the assessment process)); for candidates the crucial issue is that the 'specificity' should be high (i.e. few false positives) (Campbell et al. 1995). The most straightforward to address is the issue of false positives. Any trainee whose performance is identified as being below the minimum standard required would need to have their report form reviewed. This review should consider not only that the completion of the report had been procedurally correct (to ensure that judgements had been made on appropriate forms of evidence and to ensure that adequate warning of failure had been given), but also that the judgement made is likely to have been correct.

Of considerably greater concern is the need to minimise false negatives. Evidence provided in study two and supported from elsewhere (Campbell and Murray, 1996) strongly suggests that general practitioner trainers have considerable difficulty in

accepting that a trainee is not yet ready for independent practice; indeed in the pilot study in the west of Scotland a number of trainers refused to deny the necessary signature despite being confronted with video-taped recordings of consultations that demonstrated consultation performance below minimum standards (Campbell et al. 1993). It is therefore crucial that an attempt is made to minimise the risk of false negatives.

One approach, proposed for the consultation analysis and written submission components of this summative assessment process, is to undertake reviews of a random selection of submissions. This approach is not easily transferable for a trainer's report; it would not be acceptable to enforce a sample of trainees to have an extension of training simply to allow completion of a further trainer's report. There seem to be two realistic alternatives. One option is to use interviews with those involved in the training of a random sample of doctors who had successfully completed summative assessment. These interviews would inquire to see if there had been any reservations about the competence of the trainee by the end of the training year; this might include interviews with trainers, primary health care team workers, and course organisers. The difficulty with this approach is that it is very resource-intensive. A second option, either to be used alone or in conjunction with the approach described above, would be to attempt to monitor closely those trainees about whom reservations had been expressed during training. This might include both those trainees whose poor performance had been highlighted by their trainers earlier in the training year and those trainees whose performance, whilst not causing concern to their trainers, had caused concern to others involved in training (particularly course organisers). The principal advantage of this approach is that it is likely to be less resource intensive, but it does have the major

disadvantage that those trainees who have a problem with performance which is only apparent within the practice context (i.e. a focal problem) and which remains unrecognised by the trainer (either consciously or subconsciously) may not be identified. The other disadvantage of this approach to quality assurance is that trainees whose performance has required close scrutiny may feel hounded, and may complain that their performance is being affected by the level of scrutiny (the "observation effect" (Rowntree, 1977)).

My view is that, because of the importance of maintaining public confidence in the system of self-regulation, some form of review of the decisions that have been made will be required. On balance, probably the best choice for a system of quality assurance would be to develop a programme of interviews using a sampling approach (option one above).

6.6 Summary

The research component of this thesis has demonstrated, within the methodological limitations identified, that it is possible to develop *de novo* a trainer's report for use within a summative assessment process that can be shown to have acceptable levels of the components of validity, reliability and feasibility considered most important. Nevertheless, there are a number of areas in which further research would be justified, there is a need for a robust quality assurance system to be developed, and continuing developmental work is necessary.

This concludes the research component of this thesis. In the final chapter of the thesis a return is made to broader issues concerning assessment - in particular in relation to

summative assessment for general practice, to assessment within education, and to the place of trainer-based assessment instruments.

CHAPTER SEVEN - DISCUSSION

Chapters three to six of this thesis have concentrated entirely on the specific issue of a trainer's report as a component of a summative assessment process for general medical practice in the United Kingdom In chapter six the conclusion has been reached that a trainer's report that fulfills, to an acceptable degree, the desired technical requirements of assessment instruments has been developed.

In this final chapter a return is made to more general issues. The first section returns to the issues considered in chapter two. This is then followed by a consideration of the extent to which the work of this thesis provides insights into assessment issues. The chapter examines three questions.

Is this summative assessment process likely to address the concerns which drove its introduction? The introduction of a summative assessment process in general practice appears to have resulted from a number of driving forces (p.22-7) which are unanswered by the current regulatory process (p.33; Dunn, 1998). This section will examine whether or not it is likely that the proposed summative assessment process would address these concerns adequately.

How does this summative assessment process help us in our thinking about assessment within education? This section will consider the lessons that have emerged about assessment from the development of this particular summative assessment process. In particular it deals with the implications of applying a national entry requirement for a professional group, the implications that arise from the way in which the content of the

process has been defined, the dilemmas that arise from assessment, and the potential effect of trainer-based assessment on the trainer-trainee relationship.

How does this trainer's report help us in our thinking about assessment instruments?

The principal focus of this thesis has been a trainer's report as an assessment instrument. In this section the generalisability of application of such a report form is considered - in particular whether there are settings in which such a report form might prove valuable, whether it should be used alone as an assessment instrument, and what general lessons about the development of assessment instruments have been learnt.

7.1 Is this summative assessment process likely to address the concerns which forced its introduction?

In chapter two a number of forces driving the introduction of a summative assessment process were identified - political, societal, educational, international and professional forces. In addition, a number of arguments against summative assessment were identified. To what extent have these issues been addressed by the summative assessment process proposed?

The issues for politicians are those of ensuring adequate quality, and ensuring clear public accountability (p.22). It seems possible that a summative assessment process will go some way to reassuring politicians about the quality of doctors at the start of their careers (indeed summative assessment has now been enshrined in a regulatory framework (Anonymous, 1997)), but problems remain. In particular the continuing quality of doctors is not being assured. Whilst there may now be a much more clear level of control over entry to the profession, the absence of any formal continuing process to

ensure the maintenance of professional standards (recertification or revalidation (Parboosingh, 1998; Bashook and Parboosingh, 1998)) must surely bring into question the determination of the profession to regulate itself. This is further emphasised by the absence of any predictive validity data on the summative assessment process - surely the profession can not assume, particularly in the absence of any predictive data, that assurance of minimum standards at the point of entry to the profession will ensure that doctors' skills remain adequate for the rest of their professional lives? I believe that the significance of the development of summative assessment will be seriously undermined in the absence of any continuing assurance of acceptable levels of performance of general practitioners. Some form of recertification is needed as a matter of urgency. One option for recertification might appear to be to roll out the current summative assessment process on a regular basis. If this were to be done I do not believe it would currently be possible to include a trainer's report in such a system. This is because the completion of such a report requires a long-term relationship, the nature of that relationship being that ultimately one member (the trainer) has the power to prevent the other from holding the necessary licence. No such authority is currently invested in any one established practitioner over other practitioners.

Society is concerned about the skills of doctors serving it (p.23-4). Profound concerns about professional self-regulation as a mechanism for protecting the public from unskilled doctors have come to the surface during the second half of the nineteen-nineties (Pook and Copley, 1998); summative assessment is likely to answer these concerns to only a very limited degree. I believe that self-regulation does have some strengths (p.29) and that, with considerable modification to ensure both public involvement and continuing (and regular) assurance of standards, self-regulation could continue to fulfill

the needs of the public. It is regrettable that the opportunity to address some of these issues, particularly the involvement of the public, in the development of this summative assessment process has been missed.

For the educationalists, the proposed summative assessment process does fulfill the need for a robust assessment process as part of an educational process, particularly as the content has been specifically aimed to reflect the needs of the service for which the training is preparing these doctors. The vocational training system, described as a strong educational process on p.30, is now supported by a more rigorous assessment process which can enable those who are fit (or unfit) for the purpose to be selected (or deselected). Nevertheless, although the use of a vocational model for the selection of contents may reduce the direct influence of assessment on the curriculum, it seems possible that training will be affected in some way. Research is needed to understand that effect so that attempts can be made to ensure that the effects on training are beneficial (e.g. by supporting a particular balance of content).

The development of a rigorous end-point assessment supports the credibility of the JCPTGP as a "competent authority" for certification under international legislation (Anonymous, 1993); this legislation enables the free flow of suitably qualified practitioners within the European Economic Area. However, if the public are to be protected from all doctors who perform poorly, it will be important that the equivalence of certification processes of other member states is examined.

As far as professional forces are concerned, the research work of this thesis has certainly provided evidence that confirms the need for a change in the mechanism by which

doctors who are completing training in the United Kingdom for independent general medical practice are certificated. The work of this research programme and others (Campbell et al. 1995; Lough et al. 1995) means that a robust process of assessment as a basis for certification can be developed. The views of the profession on format, content and standards have been included in the development of this trainer's report.

The arguments against summative assessment

On p.33, based on the work of Rowntree (Rowntree, 1977), five risks of assessment were identified as being of particular significance in the setting of summative assessment the purpose of which is selection/deselection. These are the risks that arise from curriculum backwash effects, deselection effects, stereotyping, observation effects, and bureaucracy.

Curriculum backwash effects and deselection effects are both facets of the impact of the assessment process on the training curriculum. It has already been suggested (p.62) that, despite the use of a vocational model to determine the content of the assessment process, it is possible that there will be some effect on the curriculum of training and further research in this area is warranted. There must be a real risk that significant energy will be deflected away from learning to be an effective practitioner into attempting to ensure expertise in the techniques required for passing the assessment instruments. Unless such techniques are relevant to patient care, such an emphasis, however inevitable, is undesirable.

The proposed inclusion, within the process, of instruments which involve assessment by assessors who are not familiar with the trainee does reduce the risk of stereotyping (that

is, that the assessor's judgement is shaped by initial opinion), a risk that is particularly likely with an instrument such as the trainer's report.

Observation effects (the degree to which the performance of assessees will be affected by the knowledge that their performance is being observed) are of particular importance for performance tests. It is difficult to predict how great they will be. It is very likely that observation effects will occur with the assessment of videotaped consultations. Conversely, for the trainer's report, whilst at first sight observation effects seem highly likely, I am unsure that trainees will be able to adapt their performance through a whole training year; it seems possible that long-term observation may offer a greater chance to observe true performance than the shorter observation periods needed for the other instruments.

The risk of trainees failing the process purely as a result of the bureaucracy of the process must present a significant risk in a process which involves four components, three of which would require documentation to be sent to external assessors (Vocational Training Summative Assessment Board, 1998). Of particular concern is the risk that technical problems with videotape recording will result in some trainees having to re-record tapes, thereby introducing a significant time delay into the process which may then result in a delay in the doctor receiving certification. Whilst the introduction of a national protocol (Vocational Training Summative Assessment Board, 1998) might go some way to ensuring that these risks apply to all trainees equally this risk remains significant.

Summary

There are some strengths to this summative assessment process. It provides some assurance of the quality of doctors entering independent general medical practice; it provides an robust assessment process to support a complex training process; and it retains a significant professional input to the regulatory process.

These strengths are certainly balanced, and possibly outweighed, by some significant problems. The absence of either predictive validity data or some move towards a recertification process mean that summative assessment can only be considered to provide limited protection to the public. The absence of any form of public involvement in the process maintains the distance between the profession and society and may, consequently, serve to reinforce the calls for an alternative system that is more openly accountable to the public. There remain the risks of negative effects on training resulting from the process and of doctors failing purely as a result of the bureaucracy of the system.

For these reasons I believe firmly that the proposed summative assessment process in its current form must be considered only to provide a first step towards fulfilling the requirements of those groups with an interest in the quality of general practitioners in the U.K. However well researched the components may be, the process cannot be considered to be any more than an initial step in the protection of the public. Considerable further work is needed.

7.2 How does this summative assessment process help us in our thinking about assessment within education?

This section considers the lessons that have emerged from the development of this particular summative assessment process that might apply in other settings. Four issues are considered: the implications of applying a national entry requirement to a professional group; implications arising from the way in which the content of this process has been defined; dilemmas that arise within an assessment process; and the effect of a summative assessment process on the trainer-trainee relationship.

Applying a national assessment process to a professional group

The purpose of summative assessment for general practice is the selection of those doctors who, at the completion of their training, are suitable for independent general practice; it establishes a national entry requirement for independent practice. Within medicine an assessment process with the purpose of regulation at the completion of specialist training is unique. Three questions about the application of similar processes elsewhere arise: is regulation at the completion of specialist training required, by whom should it be managed, and if such an approach is to be applied elsewhere what particular requirements should be met by the assessment process?

Society must be able to have confidence in the professional groups that serve it; to do so, there must be some sort of process that enables certification of those with specialist skills.

However, as the skills of an individual professional become more and more specialised, particularly when a group of professionals with new skills is established (e.g. liver

transplant surgery), the body of professionals with similar skills becomes smaller and smaller. I believe that national certification processes apply well when the body of peers remains substantial (e.g. for general practitioners, osteopaths, or solicitors), but that for the highly specialised professional (e.g. the interventional radiologist or the tax-specialist barrister) specialist certification will have to be based on a review of the skills of that specialist by equally specialised peers. The certification process moves from a national process to a delegated process depending on the degree of specialisation.

Irrespective of the ultimate nature of the process I believe that all certification processes should share some qualities. They should be transparent - the content of the assessment, and the criteria against which the judgement will be made, should be available to assessor, assessee and the general public. They should be fair - there should be some form of quality assurance mechanisms that aim to minimise the number of incorrect judgements. They should be consistent - the standards used should be the same for all of those assessed.

Implications of the selection of content

One of the identified strengths of the work of this thesis is the method of selection of content for the summative assessment process. The strength arises firstly because the contents have been directly based on the ultimate work of the independent practitioner, and secondly because the selection of attributes from a large possible range has been based on evidence collected in a systematic way from a large body of the profession.

A number of lessons applicable to the selection of content for assessment processes arise from the work of this thesis. Firstly, the method used for the selection of contents of an

assessment process must depend on the purpose of that process; I would not advocate the same approach to the setting of the content for all assessment processes or instruments. This approach seems to have particular merit, and is consequently directly applicable, when the assessment process serves the express purpose of selection for a vocation with defined content; it would not be applicable when the express function of the assessment process lay elsewhere - for example a process designed to assess the general academic ability of a student. Secondly, a clear definition of the total content of the assessment process should be the next step to follow the declaration of the purpose of the assessment process, and the selection or development of instruments should follow (and not precede) this definition of the total content to be assessed. Thirdly, the method used in this thesis depended on the existence of a comprehensive, and widely accepted, description of the work of the profession; if no such work exists, a different approach would be needed. Fourthly, the indirect relationship between the contents of assessment processes and the contents of training curricula must be emphasised to all those involved in training. Many may assume that the relationship is direct, the consequence being that their training curriculum will be skewed by their perceptions of the relationship. In particular, when a vocational model for content selection is used, it must be emphasised that both teaching and assessment have a single purpose - namely to produce individual practitioners fit for the needs of the service for which they are training; if educators and trainees understand that, although the relationship is indirect, if the methods used for teaching and learning are targeted at ensuring training which aims to fulfill the needs of the service, passing the assessment process should become a by-product of the training rather than an end in itself.

Assessment dilemmas

In the development of this summative assessment process, and a trainer's report in particular, a number of dilemmas have become apparent. It is likely that these dilemmas would exist wherever an assessment process, particularly one that includes trainer-based assessment, is applied.

Two dilemmas have already been considered:

- **the dilemma of purpose:** should the aim of assessment be the assurance of minimum standards or the selection of those with high academic potential? This dilemma was considered on p.22. My view is that it is difficult to design an assessment instrument that will contemporaneously test at both of these levels; the instrument should be based on the purpose of the assessment process. If more than one purpose is sought, more than one instrument will be needed.
- **the dilemma of ownership:** who owns the process? For many assessment processes ownership rests with the educators; for this process the ownership lies with the profession. It has already been suggested (p.218) that one weakness with the summative assessment process proposed for general practice is the lack of involvement of those who will ultimately receive the service the quality of which is supposed to be assured by the process - should the process not ultimately be owned by the public?

The other dilemmas are:

- **the dilemma of measurement:** in the selection of attributes to be assessed, which should carry greater weight - importance or measurability? The dilemma, which has

been recognised elsewhere (Tonesk, 1983; Eliot, 1991; Wolf, 1995) concerns the tension that exists between the 'measurable-but-meaningless' and the 'important-but-impossible'. My view is that importance should take preference; if measurability dominates, the test is unlikely to have content or predictive validity.

- **the dilemma of completeness:** in the overall contents of the assessment process, what balance should be achieved between the attempt to cover all important areas and maintaining a feasible instrument - should the process be 'complete-but-complex' or 'simple-but-simplistic'? I believe that the answer to this dilemma probably lies in the importance placed upon the outcome of the assessment. If this is a high-stakes assessment, then completeness should predominate over complexity; for a low-stakes assessment process, it may be more acceptable for the feasibility of the test to predominate.
- **the dilemma of roles:** when assessment is delegated to those who are teaching, there is a risk that confusion may arise as to whether, at any particular time, the relationship is one of teacher-learner or one of assessor-assessee. I believe that this risk is balanced by the advantage of the considerable knowledge that can be gained from a long-term relationship between the two, and that the risk can be minimised by making explicit when each role is being adopted.
- **the dilemma of relationship:** this occurs when assessor and assessee are already known to each other. Again, should the advantages of assessment based on knowledge gained from a long-term relationship be lost in the desire to reduce collusion between assessor and assessee? I believe that trainers do have a unique contribution to make to assessment but that all processes that include a trainer-based assessment will need either to find ways of minimising collusion, or to ensure that

trainer-based assessment is balanced by other relationship-independent assessment instruments.

- **the dilemma of individuality:** this dilemma arises when instruments are used that require judgements to be made by a single observer. Should the simplicity and flexibility offered by a simple one-to-one assessment be replaced by the complexity of a multi-observer assessment in order to minimise bias? Again, I believe that if the system is to be (and seen to be) fair, when single-assessor instruments are used they will require balancing with assessments undertaken by other assessors, regular calibration of assessors, or a quality assurance system that enables review of assessments made.

Undoubtedly these dilemmas present considerable problems for those designing assessment processes. Whilst I have indicated my own views as to how each of these dilemmas might be managed, ultimately those designing assessment processes will have to make judgements as to the best balance in each of these areas in order to ensure that the declared purpose of the assessment process is best fulfilled. There is no simple, uniformly correct, answer to these dilemmas.

The effect on the trainer-trainee relationship

It can be seen from the arguments above that three dilemmas will be of particular significance when the process includes an instrument that requires the trainer to assess the trainee - the dilemmas of role, relationship and individuality. In addition, any summative assessment process that includes assessment of the trainee by the trainer also contains the risk of affecting the relationship between the two. In particular there is the potential risk that the trainer will be perceived as having yet more power over the

trainee. Whilst this may be a risk when introducing trainer-based assessment instruments for the first time, I believe that this is not likely to be a major issue in general practice in the U.K. - if both trainer and trainee are aware of the need for trainer-based assessment from the beginning of the relationship, and are both aware of the contents and standards required, the replacement of previously unclear assessment with an assessment process that is clear to both parties may make the relationship more honest.

A particular risk that does remain is that continuous assessment in such a long-term one-to-one training setting may alter the educational potential of the training relationship - in particular that the time (and energy) required for assessment will be undertaken at the expense of training. This is an aspect of the curriculum effects and research to consider this issue would be valuable (p.206).

Are the gains from including assessment by a trainer so great that the risks identified above are justified, or should such instruments simply be replaced with those that use external assessors rather than the trainers themselves? I believe that particular support for the use of trainer-based instruments arises when it is desired to assess performance. Firstly, trainers can observe the trainee in their usual work with considerably less interference than the imposition of external assessors. Secondly, the trainer is able to take into account the exact context in which the individual is working; if external assessors were to be used to assess performance their lack of intimate knowledge of the setting might prejudice their judgement of the performance that they observe. Thirdly, trainers can collect evidence in a much more continuous way; at best external assessors will undertake intermittent collection of evidence.

In summary I believe that trainer-based assessment has a particular role when performance is to be assessed; the advantages are then likely to outweigh the risks that arise with trainer-based assessment. Conversely, external assessors have a particular role when competence is being assessed - usual performance is no longer being sought and the risks associated with trainer-based assessment can be removed.

Summary

Experience gained in the development of a summative assessment process for general practice could be applied elsewhere. A number of lessons are particularly important.

National certification processes, however managed, should be fair, consistent and transparent.

When designing an assessment process, the purpose of the process must always come first. The contents can then be defined. The selection, or development, of instruments should only occur after these first two stages. It is possible to model the content of the process directly on the needs of the service for which trainees are being trained, but this should only be applied when the purpose of the assessment process is also directly related the needs of the service.

Dilemmas will arise in the development of an assessment process. Although guidance as to how these dilemmas might be considered can be offered, ultimately those designing the process will have to make judgements about how the process should deal with each of these dilemmas. Although some of these dilemmas arise particularly when trainers are

used as assessors for their trainees, trainers do have a particular role when the assessment of performance is desired.

7.3 How does this trainer's report help us in our thinking about assessment instruments?

In chapter three (p.53), and again in the previous section, it has been argued that the rationale for including a trainer's report as an instrument within an assessment process is that it enables assessment to be made on the performance of the trainee based on evidence collected over a prolonged period of time during which the trainee performs in the professional setting typical of that for which they are training. It has been concluded in chapter six that it is possible to develop a trainer's report that offers the opportunity to use both psychometric and impressionistic approaches to the assessment of performance and that can enable criterion-referenced measurement. In this section three issues are considered: in what particular settings is a trainer's report of this nature most suitable; should it be used alone; and what general lessons about the development of assessment instruments have arisen?

In what particular settings is a trainer's report of this nature most suitable? It has been argued in the previous section that trainers have a role in assessment, particular when assessment of performance is desired. I believe that a trainer's report is of particular value when the aspects of performance under assessment involve complex interactions of knowledge and skill (e.g. in the judgement as to whether or not to refer a patient for an investigative procedure) or when attitudes are to be scrutinised. The strength of the trainer's report is that, when judging the evidence available, the trainer can use both their understanding of the complexities involved in such actions, and their understanding of

the context in which the action was undertaken. This is likely to be particularly suitable when the requirement is the assessment of a complex attribute within the setting of training for a particular purpose. It is therefore likely to be of particular benefit in assessment processes associated with vocational models of training. When assessment of a pure attribute is required (e.g. the assessment of a particular aspect of knowledge or a particular motor skill), which is particularly likely to occur in the assessment of competence, the issue of complexity is reduced and a trainer's report is less likely to be of particular benefit. Similarly, when complex attributes are to be assessed in the setting of generic academic development, the importance of context is reduced and, again, a trainer's report is less likely to be of particular benefit.

Should the trainer's report be the sole instrument within an assessment process? Whilst an instrument that enables assessment based on evidence accrued over a significant period of time may look attractive, I believe that it should rarely be used alone. Firstly, when the assessment process is a high-stakes process, I believe that the risks are too great to defer the whole assessment to the judgement of a single individual. Secondly, when attributes of high importance are to be assessed, unless the trainer's report can be demonstrated to enable assessment of those attributes with high levels of reliability, it is useful to use at least one other instrument to provide cross-referencing (a process often referred to as triangulation). Thirdly, it has been argued (p.44) that competence tests (particularly those which break complex attributes into their component parts) may be a particularly useful addition when difficulties arise as they enable a diagnostic approach to be taken.

Finally, what general lessons about the development of new assessment instruments have arisen from this work? I believe that three particular lessons are apparent.

Firstly, in the development of assessment instruments, the strength of political arguments may equal (and sometimes be greater than) academic arguments. In this work the studies examining the contents and standards and the field test were all substantially affected by political issues.

Secondly, that approaches that involve those who will ultimately be intimately involved with the assessment can be used in the development of an instrument to promote feasibility and overall validity.

Thirdly, that the credibility of the instrument may be undermined if all stakeholders are not consulted (including service users for instruments designed to support vocational models of training).

Summary

Trainers' reports are likely to be of particular value when complex attributes are to be assessed as part of a vocational model of training. They are less likely to be useful either when competence is to be assessed or when the training is geared towards general academic development. Trainer's reports should rarely be used as the only assessment method.

In the development of new assessment instruments account must be taken of the force of political arguments as well as academic arguments; the views of those intimately involved

in the assessment are likely to prove useful in promoting feasibility and validity, and the involvement of all stakeholders is required.

7.4 Conclusions

The current regulation of entry to independent general medical practice in the U.K. is unsatisfactory. A process of summative assessment which includes performance tests offers a solution. Whilst in many ways its introduction marks a willingness to move towards the assurance of the quality of general practitioners, its introduction must only be considered a start in this process.

A trainer's report for use within such a process has been developed and tested - a new instrument in this setting, and one which fulfills the requirements of assessment instruments as fully as possible. Although evidence about its development and its properties has been presented, limitations remain and further research is justified.

The nineteen-eighties and nineteen-nineties have seen a tremendous drive towards accountability of the professions to the public. This has made the issue of ensuring high standards of professional performance a high profile issue in many arenas. By whatever means it occurs, the continued development of improved assessment techniques in the education of general practitioners should be encouraged by all members of the profession. As Lowell suggested:

"New occasions teach new duties: time makes ancient good uncouth;
They must upward still, and onward, who would keep abreast of truth."
James Russell Lowell, in The Present Crisis (1845).

BIBLIOGRAPHY

The layout of this thesis is based on the guidelines published by the University of Warwick in the booklet "Guide to Examinations for Higher Degrees by Research" (September 1996). References are quoted in the Harvard style.

Title iv, Council Directive 93/16/EEC. (1993).

Medical Act, 1858. (1858)

Medical Act, 1886. (1886) C. 28.

Medical Act, 1978. (1978) C. 12.

Medical Act, 1983. (1983) C. 54.

National Health Service (Vocational Training) Regulations 1979. (1979).

NHS (Vocational Training for General Medical Practice) Regulations. (1997).

Anonymous (1989) *Oxford English Dictionary.* Oxford: Clarendon Press.

Anonymous (1994) *The certification and recertification of doctors.* Cambridge: Cambridge University Press.

Akhurst, C (1978) Assessment of performance in professional practice in social work courses. *Assess. High. Educ.* 4 46-59.

Angoff, W.H. (1971) Scales, norms and equivalent scores. In: Thorndike, R.L (Ed.) *Educational measurement.* pp. 508-600. Washington, DC: American Council on Education.

Armstrong, J.S. and Lusk, E.J. (1987) Return postage in mail surveys: A meta-analysis. *Public Opinion Quarterly* 51 233-248.

Asher, J.J and Sciarrino, J.A. (1974) Realistic work sample tests: a review. *Personnel Psychology* 27 519-533.

Attwood, M. Johnson, N. (personal correspondence). 1998. National Summative Assessment Results - Second Year.

Baker, R. (1996) Characteristics of practices, general practitioners and patients related to levels of patients' satisfaction with consultations. *Br. J. Gen. Pract.* 46 601-605.

Baker, R. and Streatfield, J. (1995) What type of general practice do patients prefer? Exploration of practice characteristics influencing patient satisfaction. *Br. J. Gen. Pract.* 45 654-659.

Barker, D (1993) The Management Charter Initiative: an interim assessment. *Assessment and Evaluation in Higher Education* 18 125-134.

Bashook, P. and Parboosingh, J. (1998) Recertification and the maintenance of competence. *BMJ* **316** 545-548.

Bean, K.L. (1953) *Construction of educational and personnel tests.* New York: McGraw Hill.

Berwick, D.M. (1992a) Quality management in the NHS: the doctor's role - I. *BMJ* **304** 235-239.

Berwick, D.M. (1992b) Quality management in the NHS: the doctor's role - II. *BMJ* **304** 304-308.

Black, H.D. and Devine, M.C. (1986) *Assessment purposes. A study of the relationship between diagnostic assessment and summative assessment for certification.* Edinburgh: Scottish Council for Research in Education.

Black, P.J. (1993) Formative and summative assessment by teachers. *Studies in Science Education* **21** 49-97.

Bland, M. (1987) *An introduction to medical statistics.* Oxford: Oxford University Press.

Bloom, B.S. (1956) *Taxonomy of educational objectives: book 1: cognitive domain.* London: Longman.

Bloom, B.S., Krathwohl, D.R. and Masia, B.B. (1964) *Taxonomy of educational objectives: book 2: affective domain.* New York: David McKay.

Bloom, B.S. (1968) Toward a theory of testing which includes measurement-evaluation-assessment. Los Angeles, California: Center for the Study of Evaluation of Instructional Programs.

Bloom, B.S., Hastings, J.T. and Madeus, G.F. (1971) *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill.

Bowmer, I. (1994) Standard setting in certification tests. In: Newble, D., Jolly, B. and Wakeford, R. (Eds.) *The certification and recertification of doctors.* pp. 126-133. Cambridge: Cambridge University Press.

Brearley, S. (1996) Seriously deficient professional performance (editorial). *BMJ* **312** 1180-1181.

Brennan, P. and Silman, A. (1992) Statistical methods for assessing observer variability in clinical measures. *BMJ* **304** 1491-1494.

Broadfoot, P. (1979) *Assessment, schools and society.* London: Methuen.

Calnan, M., Katsouyiannopolous, V., Ovcharov, V.K., Prokhorskas, R., Ramic, H. and Williams, S. (1994) Major determinants of consumer satisfaction with primary care in different health systems. *Fam. Pract.* **11** 468-478.

Campbell, L.M., Howie, J.G.R. and Murray, T.S. (1993) Summative assessment: a pilot project in the west of Scotland. *Br. J. Gen. Pract.* **43** 430-434.

Campbell, L.M., Howie, J.G.R. and Murray, T.S. (1995a) Use of videotaped consultation in summative assessment of trainees in general practice. *Br. J. Gen. Pract.* **45** 137-141.

Campbell, L.M. and Murray, T.S. (1996) Summative assessment of vocational trainees: results of a 3-year study. *Br. J. Gen. Pract.* **46** 411-414.

Cangelosi, J.S. (1990) *Designing tests for evaluating student achievement.* White Plains, New York: Longman.

Carline, J.D., Wenrich, M. and Ramsey, P.G. (1989) Characteristics of ratings of physician competence by professional associatees. *Evaluation and the Health Professions* **12** 409-423.

Carney, T. (1992) A national standard for entry into general practice. *BMJ* **305** 1449-1450.

Cartwright, A. (1967) *Patients and their doctors. A study of general practice.* London: Routledge and Kegan Paul.

Cartwright, A. and Anderson, R. (1981) *General practice revisited. A second study of patients and their doctors.* London: Tavistock Publications.

Centre for Primary Care Research, University of Manchester. (1988) Rating scales for vocational training in general practice. Occasional Paper 40. London: Royal College of General Practitioners.

Chauncey, H. and Dobbin, D.E. (1963) *Testing: its place in education today.* New York: Harper and Row.

Chomsky, N. (1965) *Aspects of the theory of syntax.* Massachusetts: MIT Press.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* **20** 37-46.

College of General Practitioners. (1966) Evidence of the College to the Royal Commission on Medical Education. London: Royal College of General Practitioners.

Couch, A. and Keniston, K. (1960) Yeasayers and naysayers: Agreeing response sets as a personality variable. *Journal of Abnormal and Social Psychology* **60** 151-174.

Cox, J. and Mulholland, H. (1993) An instrument for assessment of videotapes of general practitioners' performance. *BMJ* **306** 1043-1046.

Crocker, L. and Algina, J. (1986) *Introduction to classical and modern test theory.* New York: Holt, Reinhart and Winston.

Cronbach, L.J. (1964) *Essentials of psychological testing.* 2nd edn. New York: Harper and Row.

Cronbach, L.J. and Meehl, P.E. (1955) Construct validity in psychological tests. *Psychological Bulletin* **52** 281-302.

Cruess, S.R. and Cruess, R.L. (1997) Professionalism must be taught. *BMJ* **315** 1674-1677.

D.E.S. (1987) Task Group on Assessment and Testing: A Report. London: Department of Education and Science and Welsh Office.

Dacre, J. Johnson, N. (personal correspondence) 1996. Standard-setting for clinical skills.

Dauphinee, D. (1994) Standard setting for recertification. In: Newble, D., Jolly, B. and Wakeford, R. (Eds.) *The certification and recertification of doctors.* pp. 201-215. Cambridge: Cambridge University Press.

Dauphinee, D., Fabb, W.E., Jolly, B., Langsley, D., Wealthall, S. and Procopis, P. (1994) Determining the content of certifying examinations. In: Newble, D., Jolly, B. and

Wakeford, R. (Eds.) *The certification and recertification of doctors.* pp. 92-104. Cambridge: Cambridge University Press.

De Groot, A.D. (1969) *Methodology: Foundations of inference and research in the behavioural sciences.* The Hague: Mouton.

Department of Health (England) (1996) Choice and opportunity (Primary Care: The Future). London: Department of Health.

Department of Health (England) (1997) Delivering the future. London: Department of Health.

Desforges, C. (1989) *Testing and assessment.* London: Cassell Education Ltd.

Difford, F. and Hughes, R.C.W. (1992) Rating scales for the assessment of vocational trainees (letter). *Br. J. Gen. Pract.* **42** 79.

Dillman, D.A. (1978) *Mail and telephone surveys: The total design method.* New York: Wiley.

Doyle, C. (1996) Questions patients should ask. *The Daily Telegraph* 9 November edn.

Dunn, P.M. (1998) The Wisheart affair: paediatric cardiological services in Bristol, 1990-5. *BMJ* **317** 1144-1145.

Dwyer, C.A. (1990) Trends in the assessment of teaching and learning: educational and methodological perspectives. In: Broadfoot, P., Murphy, R. and Torrance, H. (Eds.) *Changing educational assessment. International perspectives and trends.* London: Routledge.

Ebel, R.L. (1979) *Essentials of educational measurement.* Englewood Cliffs, New Jersey: Prentice Hall.

Eliot, J. (1991) *Action research for educational change.* Oxford: Oxford University Press.

Fabb, W.E. and Marshall, J.R. (1983) *The assessment of clinical competence in general family practice.* Lancaster: MTP Press.

Farmer, A. (1991) Setting up consensus standards for the care of patients in general practice (editorial). *Br. J. Gen. Pract.* **41** 135-136.

Flanagan, J.C. (1954) The critical incident technique. *Psychol Bull* **51** 327-358.

Fletcher, D. (1996) Dorrell bolsters the 'key' fole of family doctors. *The Daily Telegraph* 18th December edn.

Frey, J.H. and Fontana, A. (1991) The group interview in social research. *Social Science Journal* **28** 174-187.

Gallagher, M., Hares, T., Spencer, J. and Bradshaw, C. (1993) The nominal group technique: a research tool for general practice? *Fam. Pract.* **10** 76-81.

Confidence interval analysis version 1.2. Gardner, M.J., Gardner, S.B. and Winter, P.D. (1992) 1.2. London: BMJ.

General Medical Council 1993. The Medical Register. London: GMC.

General Medical Council 1994. The Medical Register. London: GMC.

Gipps, C. and Stobart, G. (1993) *Assessment: a teacher's guide to the issues.* London: Hodder and Stoughton.

Glaser, E.M. (1980) Using behavioural science strategies for defining the state-of-the-art. *J Appl Behav Sci* **16** 79-92.

Glaser, R. (1963) Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist* **18** 519-521.

Glaser, R. (1971) A criterion-referenced test. In: Popham, W.J. (Ed.) *Criterion-referenced measurement.* Englewood Cliffs, New Jersey: Prentice Hall.

Gold, J.A. (1981) Wiser than the laws?: the legal accountability of the medical profession. *Am. J. Law Med.* 7 145-181.

Gonczi, A. (1994) Competency based assessment in the professions in Australia. *Assessment in Education* 1 27-44.

Government Statistical Service (1996) Statistics for General Medical Practitioners in England: 1985-1995. Bulletin 1996/6, London: Department of Health.

Graham, D.T. (1981) The training and examination of radiographers - a plea for continuous assessment. *J. Furth. High. Educ. Scotl.* 6 26-38.

Gray, D.J.P. (1977) A system of training for general practice. Occasional paper 4. London: Journal of the Royal College of General Practitioners.

Gray, D.P. (1992) The story of post graduate education. In: Gray, D.P. (Ed.) *Forty years on. The story of the first forty years of The Royal College of General Practitioners.* pp. 109-117. London: Atalink.

Guilford, J.P. (1954) *Psychometric methods.* New York: McGraw-Hill.

Haile, P.J. (1977) An assessment typology. *Social Policy* 8 20-26.

Harden, R.M. (1979) Assess clinical competence - an overview. *Medical Teacher* 1 289-296.

Hart, I.R. (1992) Trends in clinical assessment. In: Harden, R.M., Hart, I.R. and Mulholland, H. (Eds.) *Approaches to the Assessment of Clinical Competence. Part 1.* pp. 17-26. Norwich: Page Brothers.

Haslam, D. (1998) MRCGP (letter). *Br. J. Gen. Pract.* **48** 1003.

Hasler, J. Johnson, N. (personal correspondence) 1994. Development of trainer's reports.

Hibbs, J. (1995) Bad doctors face exposure by colleagues. *Daily Telegraph* 7 August 1995 edn.

Hickox, M. (1995) Situating vocationalism. *British Journal of Sociology of Education* **16** 153-163.

Hudson, B. (1973) *Assessment techniques: an introduction.* London: Methuen.

Irvine, D. (1997) The performance of doctors. I: Professionalism and self regulation in a changing world. *BMJ* **314** 1540-1542.

Irvine, D.H., Pereira Gray, D.J. and Bogle, I.G. (1990) Vocational training: the meaning of 'satisfactory completion' (letter). *Br. J. Gen. Pract.* **40** 434

Johnson, N., Hasler, J., Mant, D., Randall, T., Jones, L. and Yudkin, P. (1993) General practice careers: changing experience of men and women vocational trainees between 1974 and 1989. *Br. J. Gen. Pract.* **43** 141-145.

Johnson, T.J. (1972) *Professions and power.* London: MacMillan.

Joint Committee on Postgraduate Training for General Practice Working Party on Assessment. (1992a) The interim report of the Working Party on Assessment. London: JCPTGP.

Joint Committee on Postgraduate Training for General Practice (1992b) Report 1991-2. London: JCPTGP.

Joint Committee on Postgraduate Training for General Practice (1993) Report 1992. London: JCPTGP.

Joint Committee on Postgraduate Training for General Practice (1994) The report on the work of the Joint Committee on Postgraduate Training for General Practice 1993. London: JCPTGP.

Joint Committee on Postgraduate Training for General Practice (1995) Report on the work of the Joint Committee on Postgraduate Training for General Practice 1994. London: JCPTGP.

Kandel, I. (1936) Examinations and their substitutes in the United States. Bulletin 28. New York: Carnegie Foundation for the Advancement of Teaching.

Kane, M.T. (1982) The validity of licensure examinations. *American Psychologist* **37** 911-918.

Kane, M.T. (1992) The assessment of professional competence. *Evaluation and the Health Professions* **15** 163-182.

Kee, F. (1996) Patients' prerogatives and perceptions of benefit. *BMJ* **312** 958-960.

Kenyan, A., Friedman, M. and Benbassat, J. (1987) Reliability of global rating scales in the assessment of clinical competence of medical students. *Med. Educ.* **21** 477-481.

King, I.W. and Danks, D.J. (1986) The assessment of supervised work experience. *Business Education* 7 22-28.

Klug, B. (1974) *Pro profiles.* London: NUS Publications.

Likert, R.A. (1952) A technique for the development of attitude scales. *Educational and Psychological Measurement* **12** 313-315.

Linstone, H.A. and Turoff, M. (1975) *The Delphi method: techniques and applications.* Massachusetts: Addison-Wesley Publishing Co.

Livingston, S.A. and Zieky, M.J. (1982) *Passing scores.* Princeton: Educational Testing Service.

Lough, J.R.M., McKay, J. and Murray, T.S. (1995) Audit and summative assessment: a criterion-referenced marking schedule. *Br. J. Gen. Pract.* **45** 607-609.

Mant, D. and Yudkin, P. (1993) Collecting and analysing data. In: Lawrence, M. and Schofield, T. (Eds.) *Medical audit in primary health care.* Oxford: Oxford University Press.

Merton, R.K., Fiske, M. and Kendall, P.L. (1956) *The focused interview.* New York: Free Press.

Messick, S. (1984) The psychology of educational measurement. *Journal of Educational Measurement* **21** 215-238.

Miller, G.E. (1990) The assessment of clinical skills/competence/performance. *Acad. Med.* **65**, 63-67.

Ministry of Health and Department of Health for Scotland. (1944) Report of inter-departmental committee on medical schools. London: HMSO.

Mulholland, H. and Tombleson, P.M.J. (1990) Assessment of the general practitioner. *Br. J. Gen. Pract.* **40** 252-254.

Newble, D.I. (1988) Eight years' experience with a structured clinical examination. *Med. Educ.* **22** 200-204.

NHS Executive (1996) Developing a primary care led NHS. Leeds: NHSE.

Nitko, A.J. (1977) A model for criterion-referenced tests based on use. In: Sumner, R. and Robertson, T.S. (Eds.) *Criterion referenced measurement and criterion referenced tests: some published work reviewed.* Windsor: NFER.

Norman, G.R., Neufeld, V.R., Walsh, A., Woodward, C.A. and McConvey, G.A. (1985) Measuring physicians' performance by using simulated patients. *Journal of Medical Education* 60 925-934.

Oxford region course organisers and regional advisers group. (1985) Priority objectives for general practice vocational training. Occasional paper 30, London: Royal College of General Practitioners.

Parboosingh, J. (1998) Revalidation for doctors (editorial). *BMJ* **317** 1094-1095.

Phillips, T. (1993) *Assessing competencies in nursing and midwifery education.* London: English National Board.

Pook, S. and Copley, J. (1998) Inquiry after heart doctors are struck off. *The Daily Telegraph* 19 June edn.

Popham, W.J. and Husek, T.R. (1969) Implications of criterion-referenced measurement. *Journal of Educational Measurement* **6** 1-9.

Power, C. and Wood, R. (1987) Aspects of the competence-performance distinction: Educational, psychological and measurement issues. *Journal of Curriculum Studies* **19** 409-424.

Preece, P.F.W. (1993) The assessment of teaching practice performance. *Research in Education* **49** 23-27.

Quality and Consumers Branch (1996) Patient partnership: building a collaborative strategy. Leeds: NHSE.

Quine, S. (1985) 'Does the mode matter?': A comparison of three modes of questionnaire completion. *Community Health Studies* **9** 151-156.

Rakowski, R.T. (1990) Assessment of student performance during industrial training placements. *International Journal of Technology and Design Education* **1** 106-110.

Raven, J. (1991) *The tragic illusion: educational testing.* Unionville, New York: Trillium Press.

Resnick, L.B. and Resnick, D.P. (1992) Assessing the thinking curriculum: new tools for educational reform. In: Gifford, B. and O'Connor, M. (Eds.) *Changing assessments: alternative views of aptitude, achievement and instruction.* London: Klewer Academic Publishers.

Rethans, J.J., Sturmans, F., Drop, R., van der Vleuten, C. and Hobus, P. (1991) Does competence of general practitiones predict their performance? Comparison between examination setting and actual practice. *BMJ* **303** 1377-1380.

Rhodes, M. (1998) Evaluating the trainer's report in the summative assessment of general practice trainees: a qualitative approach. University of London. pp.1-106.

Rhodes, M. and Styles, W.M. (1995) Summative assessment: towards the trainer's report. *Education for General Practice* **6** 124-130.

Rowntree, D. (1977) *Assessing students. How shall we know them?* London: Harper & Row.

Royal College of General Practitioners (1985) Policy Statement 2: Quality in General Practice. London: RCGP.

Royal Commission on Medical Education. (1968) Report, 1965-68. London: HMSO.

Schatzman, L. and Strauss, A.L. (1973) *Field Research: Strategies for a Natural Sociology*. Englewood Cliffs, NJ: Prentice-Hall.

Schmitt, N., Gooding, R.Z., Noe, R.A. and Kirsch, M. (1984) Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology* **37** 407-422.

Schwarz, N., Knauper, B., Hippler, H., Noelle-Neumann, E. and Clark, L. (1991) Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly* **55** 570-582.

Shavelson, R.J., Webb, N.M. and Rowley, G.L. (1989) Generalizability theory. *American Psychologist* **44** 922-932.

Spady, W.G. (1988) Organising for results: the basis of authentic restructuring and reform. *Educational Leadership* **October** 4-8.

Stanley, I. and Al-Shehri, A. (1993) Reaccreditation: the why, what and how questions. *Br. J. Gen. Pract.* **43** 524-529.

Statement by a working party of the second European conference on the teaching of general practice. (1977) The work of the general practitioner. *J. R. Coll. Gen. Pract.* **27** 117.

Stewart, D.M. and Shamdasani, P.N. (1990) *Focus groups: theory and practice.* Newlay Park, California: Sage Publications.

Stocking, B. (1991) Patient's charter (editorial). *BMJ* **303** 1148-1149.

Streiner, D.L. and Norman, G.R. (1995) *Health measurement scales.* Second edn. Oxford: Oxford University Press.

Suissa, S. (1991) Binary methods for continuous outcomes: A parametric alternative. *Journal of Clinical Epidemiology* 44 241-248.

Super, D.E. (1949) *Appraising vocational fitness.* New York: Harper.

Swanson, D.B., Norman, G.R. and Linn, R.L. (1995) Performance-based assessment: lesson from the health professions. *Educational Researcher* **June/July** 5-11.

Thatcher, M. **156**, London: Hansard. (1989)

The Nature of General Medical Practice Working Party (1996) The nature of general medical practice. 27, Exeter: Royal College of General Practitioners.

Thorndike, R.M. (1997) *Measurement and evaluation in psychology and education.* 6th edn. Upper Saddle River, New Jersey: Prentice Hall.

Tonesk, X. (1983) Clinical judgment of faculties in the evaluation of clerks (editorial). *Journal of Medical Education* 58 213-214.

United Kingdom Parliament House of Commons Social Services. (1975) Report of the Committee of Inquiry into the Regulation of the Medical Profession. London: HMSO.

Van de Ven, A.H. and Delbecq, A.L. (1972) The nominal group as a research instrument for exploratory health studies. *Explor Health Studies* (March):337-342.

Van der Vleuten, C.P.M., Norman, G.R. and De Graaff, E. (1991) Pitfalls in the pursuit of objectivity: issues of reliability. *Med. Educ.* **25** 110-118.

Vincent, R. (1996) Assessment in the workplace. In: Goldstein, H. and Lewis, T. (Eds.) *Assessment: problems, developments and statistical issues. A volume of expert contributions.* pp. 231-244. Chichester: J. Wiley and Sons.

Vocational Training Summative Assessment Board (1998) *Protocol for the management of summative assessment.* London: Vocational Training Summative Assessment Board.

Ward, A.W., Stoker, H.W. and Murray-Wood, M. (1996) *Educational measurement. Origins, theories and explications.* Lanham, Maryland: University Press of America.

Ward, C. (1980) *Designing a scheme of assessment.* Cheltenham: Stanley Thomas.

Weiss, C.H. (1975) Interviewing in evaluation research. In: Struening, E.L. and Guttentag, M. (Eds.) *Handbook of evaluation research.* pp. 355-395. Beverley Hills: Sage Publications.

Wensing, R., Jung, H.P., Mainz, J., Olesen, F. and Grol, R. (1998) A systematic review of the literature on patient priorities for general practice care. Part 1: Description of the research domain. *Soc. Sci. Med.* **47** 1573-1588.

Wildt, A.R. and Mazis, A.B. (1978) Determinants of scale response: Label versus position. *Journal of Marketing Research* **15** 261-267.

Williams, S.J. and Calnan, M. (1991) Key determinants of consumer satisfaction with general practice. *Fam. Pract.* **8** 237-242.

Willmott, A. (1978) Assessment and Performance. *Oxford Review of Education* **4** 51-64.

Wolf, A. (1995) *Competence-based assessment.* Buckingham: Open University Press.

Wolf, A. (1996) Vocational assessment. In: Goldstein, H. and Lewis, T. (Eds.) *Assessment: problems, developments and statistical issues. A volume of expert contributions.* pp. 209-230. Chichester: J. Wiley and Sons.

Yu, J. and Cooper, H. (1983) A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research* **20** 36-44.

Zeitlyn, B.B. (1979) A patient's charter. *BMJ* **2** 103-104.

APPENDICES

APPENDIX 1.1

Literature search strategy

a) Search terms:

- Summative
- Evaluation
- Assessment
- Summative and assessment
- End-point
- End-point and assessment
- Assessor(s)
- Trainer(s)
- Trainer(s) and assessment
- Profession(s)
- Profession(s) and assessment
- Professional
- Professional and assessment
- Performance
- Performance and assessment and profession or professional
- Medical or doctor and assessment
- Trainer(s) report
- Supervisor(s) report

b) Search databases used

Database	**Years**
ERIC - British Education Index and British Education Theses Index	1976-97
Medline	1966-97
CINAHL	1982-97
Sociofile	1974-97
PsycLit	1974-97
SIGLE (Grey literature)	1980-97

APPENDIX 4.1.

Contents questionnaire for study two

Code No.

FOR THE "IMPORTANCE" SCORE, please indicate how important the quality is for independent general practice (using the scale 1 - 5) by putting a ring around the number that represents your view.

FOR THE "METHOD OF ASSESSMENT", please indicate which methods of assessment should be used to assess that quality. Please tick all those boxes you feel apply; if you tick 'other', please specify.

a. PATIENT CARE

1. The doctor can recognise common physical, psychological and social problems

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

2. Within the assessment the doctor includes patients' beliefs, ideas, concerns, effects and expectations

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

3. The doctor considers and follows up psychological and social factors

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

4. The doctor undertakes appropriate examination with appropriate consideration of the patients needs and feelings

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

5. The doctor is able to examine each system (e.g. cardiovascular, respiratory) and each organ (e.g. eye, ear) proficiently

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

6. The doctor is able to assess the mental state proficiently

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

7. The doctor understands the principles of problem definition (including the use of hypothesis formation and testing)

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

8. The doctor copes with the anxieties felt as a result of unstructured presentations, difficulty in reaching conclusions, and lack of continuous patient monitoring

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

9. The doctor is able to use time as a diagnostic and therapeutic tool

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify.

10. The doctor chooses appropriate management for each problem with the patient (including the care of chronic problems)

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

11. The doctor uses management plans which include effective use of other members of the team

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

12. The doctor makes effective use of the records

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

13. The doctor is able to undertake the following specific elements of examination proficiently:

a) Use of auroscope

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

b) Use of ophthalmoscope

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

c) Use of sphygmomanometer

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

d) Use of stethoscope

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

13. e) Use of patella hammer

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

f) Use of tuning fork

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

g) Vaginal examination

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

h) Use of vaginal speculum

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

i) Taking of cervical smear

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

13. j) Rectal examination

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

k) Use of proctoscope

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

l) Use of laryngoscope

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

m) Use of peak flow meter

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

n) Use of the ECG

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

14. The doctor demonstrates a broad knowledge of all aspects of the appropriate use of drugs (including actions, interactions, side effects, costs, and legal aspects)

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

15. The doctor has the knowledge and skills to deal with life events and crises

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

16. The doctor provides appropriate care and support for patients and their families

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

17. The doctor has a knowledge of available agencies and resources and the skills to refer appropriately

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

18. The doctor understands the importance of involving and educating patients

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

19. The doctor is aware of the costs of his/her activities and recognises the limits to those costs

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

20. The doctor diagnoses and manages acute emergency situations appropriately

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

21. The doctor responds appropriately to requests for urgent attendance at patients

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

22. The doctor is able to undertake the following emergency procedures

a) Give an intravenous injection

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

b) Give an intramuscular injection

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

c) Undertake basic cardio-pulmonary resuscitation

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

23. The doctor understands the principles involved in prevention in general practice (including case finding, screening, health education and monitoring of preventive activities)

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

24. The doctor has a knowledge of the systems used to identify individuals and sections of the practice population (e.g. disease registers, computerised registration data)

Importance: Fairly important 1 2 3 4 Very important 5 Crucial

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐ **Other** ☐ - please specify:

25. The doctor is able to provide effective preventive services to individual patients

Importance: Fairly important 1 2 3 4 Very important 5 Crucial

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐ **Other** ☐ - please specify:

26. The doctor is able to provide effective preventive services to the population

Importance: Fairly important 1 2 3 4 Very important 5 Crucial

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐ **Other** ☐ - please specify:

27. The doctor demonstrates an understanding of the effect of social and environmental circumstances on the patient

Importance: Fairly important 1 2 3 4 Very important 5 Crucial

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐ **Other** ☐ - please specify:

b. COMMUNICATION

28. The doctor demonstrates effective communication skills when dealing with patients

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

29. The doctor demonstrates understanding and respect for colleagues

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

30. The doctor has an understanding of the importance of meetings and discussion with colleagues

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

31. The doctor demonstrates the skills to discover the strengths and weaknesses of colleagues and their need for support

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

32. The doctor uses his/her knowledge of the practice and of patients appropriately in various contacts (e.g. practice or team meetings, telephone contacts, contacts with families)

Importance: Fairly important 1 2 3 4 Very important 5 Crucial

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐ Other ☐ - please specify:

c. ORGANISATION

33. The doctor understands his/her obligations according to the NHS contract and regulations

Importance: Fairly important 1 2 3 4 Very important 5 Crucial

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐ Other ☐ - please specify:

34. The doctor understands the importance of the need to manage a practice effectively

Importance: Fairly important 1 2 3 4 Very important 5 Crucial

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐ Other ☐ - please specify:

35. The doctor is able to monitor aspects of practice activity

Importance: Fairly important 1 2 3 4 Very important 5 Crucial

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐ Other ☐ - please specify:

36. The doctor has an understanding of the basic methods of research as applied to general practice

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

37. The doctor is able to take appropriate action when organisational problems are identified

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

38. The doctor has a knowledge of the most important sections of the NHS contract and regulations with regard to sources of income and superannuation

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

39. The doctor has a knowledge of the most important organisational aspects of practice and partnership (including agreements, accounts, building, tax)

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

40. The doctor understands medico-social legislation and the impact of this on the patient

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

41. The doctor understands the application of new technology to general practice

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

42. The doctor understands the principles of successful introduction of change and innovation

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

43. The doctor is able to manage his/her own time

Importance: Fairly important Very important Crucial
1 2 3 4 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

44. The doctor is aware of his/her own limitations, the skills of others and the ability to delegate appropriately

Importance: Fairly important **1** **2** Very important **3** **4** Crucial **5**

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

45. The doctor is able to determine and respond to the health needs of the community

Importance: Fairly important **1** **2** Very important **3** **4** Crucial **5**

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

46. The doctor knows how and where to intervene in the community of behalf of others

Importance: Fairly important **1** **2** Very important **3** **4** Crucial **5**

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

d. PROFESSIONAL VALUES

47. The doctor is aware of his/her own values, beliefs and attitudes; how they are influenced; and how they affect others

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

48. The doctor recognises the social, cultural and organisational factors that define and affect his/her work

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

49. The doctor possesses and applies ethical principles

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

50. The doctor shows tolerance, respect and flexibility when responding to the ideas of others

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

51. The doctor is willing to undergo peer review and is able to give and receive criticism

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

52. The doctor is able to maintain his/her own physical and mental health

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

53. The doctor is aware of the factors that influence the relationships between personal and professional life

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

54. The doctor is willing to accept appropriate responsibility for patients, partners, colleagues and others

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

e. PERSONAL AND PROFESSIONAL GROWTH

55. The doctor is able to identify strengths and weaknesses in his/her performance

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

56. The doctor can recognise, define and respond to change, including changing needs in patients, colleagues and the community

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

57. The doctor can define his/her own educational needs and appropriate methods of meeting those needs

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

58. The doctor is able to produce change in self and others

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: Written exam ☐ External observation ☐ Trainee project ☐ Trainer's report ☐
Other ☐ - please specify:

59. The doctor is aware of the factors that limit his/her effectiveness

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

60. The doctor is able to manage his/her own time

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

61. The doctor is able to manage and overcome the factors that limit his/her effectiveness

Importance: Fairly important 1 2 Very important 3 4 Crucial 5

Assessment: **Written exam** ☐ **External observation** ☐ **Trainee project** ☐ **Trainer's report** ☐
Other ☐ - please specify:

GENERAL INFORMATION

62. How old are you? .....

63. In what year did you qualify? .....

64. For how many years have you been a trainer? .....

65. Are you male or female? Male ☐ Female ☐

66. What is the TOTAL list size for your practice? .....

67. How many partners are there in your practice (including yourself)? .....

68. In general, how long is the trainee based in your practice? Please tick one response:

0-4 months ☐
5-8 months ☐
9-12 months ☐

69. Have you ever considered NOT signing up a trainee on form VTR1? Please tick one response:

Yes ☐ No ☐

70. If the answer to question 69 is "yes" would you be prepared to undertake a confidential interview with one researcher (NJ) to discuss what issues cause difficulty at the pass/fail interface? Please tick one response:

Yes ☐ No ☐

71. **Are there any additional questions that you would like to see included in the trainer's report?**

72. **Are there any other comments you would like to make about the trainer's report?**

THANK YOU FOR YOUR TIME IN COMPLETING THIS FORM!

APPENDIX 4.2

Questionnaire to doctors recently completing training

THE CONTENTS OF A TRAINER'S REPORT FOR SUMMATIVE ASSESSMENT
Code no: .....

PATIENT CARE - *GENERAL CLINICAL SKILLS*

1: The doctor can recognise common physical psychological and social problems

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

2: The doctor is able to examine each system and each organ proficiently

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

3: The doctor has the knowledge and skills to deal with life events and crises

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

4: The doctor demonstrates a broad knowledge of all aspects of the appropriate use of drugs (including actions, interactions, side effects, costs and legal aspects)

a) This is a piece of knowledge that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this piece of knowledge is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

5: The doctor diagnoses and manages acute emergency situations appropriately

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

PATIENT CARE: *PATIENT MANAGEMENT SKILLS*

6: Within his/her assessment the doctor includes the patients' beliefs, ideas, concerns, expectations and the effects of the problem

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

7: The doctor undertakes examination with appropriate consideration of the patients needs and feelings

a) This is an attitude that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this attitude is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

8: The doctor chooses appropriate management for each problem with the patient (including the care of chronic problems)

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

PATIENT CARE: *CLINICAL JUDGEMENT*

9: The doctor provides appropriate care and support for patients and their families

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

10: The doctor undertakes appropriate examination (including investigations)

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

11: The doctor responds appropriately to requests for urgent attendance at patients

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

COMMUNICATION SKILLS

12: The doctor demonstrates effective communication skills when dealing with patients

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

PERSONAL AND PROFESSIONAL GROWTH

13: The doctor is able to identify strengths and weaknesses in his/her performance

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

ORGANISATIONAL SKILLS

14: The doctor is aware of his/her own limitations, the skills of others, and the ability to refer or delegate appropriately

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

15: The doctor is able to manage his/her own time

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

16: The doctor understands the obligations of a general practitioner according to the NHS contract and regulations

a) This is a piece of knowledge that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this piece of knowledge is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

17. The doctor has an understanding of the basic methods of research as applied to general practice

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

PROFESSIONAL VALUES

18: The doctor possesses and applies ethical principles

a) This is an attitude that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this attitude is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

19: The doctor is able to maintain his/her own physical and mental health to a level which enables him/her to discharge the duties of a general medical practitioner

a) This is an attitude that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this attitude is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

20: The doctor is willing to accept appropriate responsibility for patients, partners, colleagues and others

a) This is an attitude that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this attitude is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

21. The doctor is able to undertake the following aspects of examination proficiently AND to interpret the findings made:

a): the mental state

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b): the auroscope

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

c): the ophthalmoscope

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree not disagree ☐ agree ☐ strongly agree ☐

d): the sphygmomanometer

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

e): the stethoscope

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

f): the peak flow meter

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

g): the vaginal examination

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

h): the vaginal speculum

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

i): the cervical smear

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

j): the rectal examination (does not include proctoscopy)

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

k): the laryngoscope

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

SPECIFIC CLINICAL SKILLS: *Emergency care*

22. The doctor is able to undertake the following techniques proficiently:

a): the doctor is able to give an intravenous injection

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b): the doctor is able to give an intramuscular or subcutaneous injection

a) This is a skill that is needed in general practice:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

b) It is reasonable that this skill is assessed by means of a trainer's report:

strongly disagree ☐ disagree ☐ neither agree nor disagree ☐ agree ☐ strongly agree ☐

Are you male or female? M/F How old are you?

Are there any comments that you would like to make about the trainer's report?

THANK YOU FOR YOUR TIME. Please return this questionnaire in the stamped addressed envelope enclosed.

APPENDIX 4.3:

Guidelines for the standards group members

GUIDELINES:

- start with standard; then consider evidence needed and sources
- standard is *minimal* - "what constitutes a fail?" - and is based on "independent general practice" (see example)
- 2nd group consider topic independently of 1st group, then consider their ideas
- 3rd group consider topic independently of others, then use views of other two groups along with their own to develop consensus
- use worksheets, passing them on to the next group named at end of session
- please be clear and concise - this will need to be usable by us and our peers

APPENDIX 4.4:

Example of standards being sought

TOPIC:

The doctor is able to use the ECG proficiently

1. *MINIMUM* STANDARD - what would constitute a fail?

a) The doctor is **unable** to place the leads correctly and record an ECG

b) The doctor is **unable** to interpret ECG changes that represent life-threatening disease - in particular:

acute myocardial infarction
complete heart block
ventricular fibrillation

2. EVIDENCE NEEDED

a) Direct observation of the recording of at least one ECG

b) Correct ECG interpretation of **at least one example of each of the three conditions listed**

3. ACCEPTABLE SOURCES OF EVIDENCE

a) Recording of ECG -

personal observation of real or simulated situation
evidence from consultant physician
evidence from practice nurse accredited for undertaking ECG recording

b) Interpretation of ECG

random case analysis
problem case analysis
simulated surgery/OSCE
direct observation of consultations
evidence from partner
evidence from consultant physician

APPENDIX 4.5:

Worksheet used by the standards groups

ITEM :

	THE STANDARD: (NB *MINIMAL*)	**THE EVIDENCE:**	
	WHAT WOULD CONSTITUTE A FAIL?	**EVIDENCE NEEDED TO DECIDE?**	**ACCEPTABLE SOURCES OF EVIDENCE?**
group			
group			
group			

APPENDIX 4.6:

Questionnaire to trainees involved in field study

INFORMATION ABOUT YOU AND ABOUT YOUR EXPERIENCE OF USING THE REPORT

Code no:

1. How old are you? .....
2. Are you male or female? .....
3. What phase within your general practice year does December 1995 represent: (e.g. if this is month 3 please indicate 3; if it is month 6, please indicate 6) .....
4. Please list the difficulties that arose as a result of the trainers in your practice completing the structured trainer's report:

APPENDIX 4.7:

Questionnaire to trainers involved in field study

INFORMATION ABOUT YOU AND ABOUT YOUR EXPERIENCE OF USING THE REPORT

1. How old are you? .....
2. For how many years have you been an approved trainer? .....
3. Are you male or female? .....
4. Please list the difficulties that arose in competing this form:

5. Please list improvements that you would like to see in the report (including the guidance notes and the layout/design)

6. Please provide any hints for future users of the report

7. Are there any other comments that you would like to make about the report?

APPENDIX 5.1

Results of interviews with trainers' groups

ONE.

I. Current system: log based on monthly assessment and formal mid-term assessment. Mid-term assessment involves trainer, trainee and external assessor with norm reference scoring system in 8 areas. Keen to avoid check lists.

II. Pros: check on knowledge/problem solving/consulting skills. External assessment.

III. Cons: mainly formative. Concerns re collusion between trainer and trainee. Curriculum driven by assessment.

IV. Suggestions for report:

- A. take trainer out of it
- B. aim for objective measurements
- C. criterion referencing
- D. humane treatment of referred trainees
- E. avoid check lists
- F. may need subjectivity in some areas, esp. attitudinal
- G. need validation with another assessor
- H. use discussion with trainers group/CO as part of evidence
- I. suggest experiment concerning self-directed learning.

V. Hassles with trainees. Yes - mainly attitudinal.

TWO.

I. Current system: formal mid-term assessment, using trainers/ trainees/external assessor. There is considerable emphasis on evidence-based reporting. Main aim is as formative assessment. Mid-term assessment currently uses a modified MRS with additions on management and technology. Also use of Critical Event monitoring. MEQ/MCQ also used. Priority objectives based.

II. Pros: external assessment. Timing allows remediation. Use of critical event monitoring particularly in relation to attitudinal problems. Breadth of assessment. Open assessment.

III. Cons: recognition of need to include management/technology.

IV. Suggestions:

A. avoid over-bureaucratisation

B. training of trainers, particularly to increase their observation of the trainee

C. emphasise the need for documentation all way through year

D. particular emphasis on attitudes in all areas

E. consider how it is going to be fitted in as an educational tool

F. support for trainer training

G. look at ways of overcoming mentor/assessor dilemma

H. summative assessment needs to be the summation of formative assessments without remediated negatives being carried over

I. need for input from CO

J. major emphasis on attitudes

V. Hassles with trainees - yes, mainly attitudinal.

THREE.

I. Current system: only system is mid-term assessment. Basically free text. Undertaken at 6 months. Two external assessors. Principally an assessment of training.

II. Pros: evidence usually written down. Group do not feel scales to be helpful.

III. Cons: trainers commented that it is often difficult to write.

IV. Suggestions for report

A. start early

B. make explicit

C. consider how selection for General Practice in the first place should be made.

D. need for sensitive approach to those who are not signed up.

E. trainer/assessor dilemma.

F. feeling that Structured Trainers Report may make signing up easier

G. suggest three group summary - pass, needs help, fail. Give evidence for all.

H. use as the final Formative Assessment

I. for criterion referencing of attitudes, suggest use of "How does it affect care of patients?"; "Is it isolated or general?".

J. involve Course Organiser

V. Hassles with trainees - yes, mainly in communications/organisation/attitudinal/energy. Were discussed within Trainers Group.

FOUR.

I. Current System: one sub-group has recently looked on a summative assessment based on priority objectives. A great number of tools used, including videos, log books, MCQ, MEQ, Oxford Check List, Practical Skills Check List, Chronic Disease Check List, Feed back from Primary Care Team, Learning Staff Questionnaire, PEP, NTA, Simulated Surgeries. Have recently looked at the possibility of using Morell Patient Satisfaction Questionnaire. Mid-term assessment currently used involves two external assessors.

II. Pros: useful as a Structured Reference.

III. Cons: lengthy. How objective? - only one person's opinion.

IV. Suggestions for Report

A. Scoring System Yes/No/Yes but.

B. Put in trainees major strengths and errors where development would help

C. Trainees property, except for Yes/No answers

D. Suggest use of summation of formative assessments with formal warnings to prevent carrying forward negative assessment

E. Need for use of evidence.

V. Hassles with trainees - yes, medical skills, communication skills, health related. Trainers report might have helped to deal with Health Questions.

FIVE

I. Present System: mid-term assessment. Mainly as an assessment of training. Trainee led. Partly based on OXVT 7.

II. Pros: assessment of Training.

III. Cons: not discussed.

IV. Suggestions for report.

A. Summation of formative assessment.

B. Need for review of those who failed by Senior Adviser

C. Needs to be short and concise - do not include desirable assets, except in part of report to be used as formative assessment if wished.

D. Use of YES/NO format, with evidence to be given. It is felt that some questions may need a more expansive reply.

E. End with comments at the end

V. Hassles with trainees - yes, relationships/personality. Report might have offered legal protection to trainers who did not wish to sign up trainees.

SIX

I. Current system: principal system is mid-term assessment, using video-taped tutorial/surgery. Basically external assessment, principally of training. The group also recognised that current trainer references is a form of summative assessment. Discussion based principally on trainer reference.

II. Pros: currently trainers tend to aim for strengths and weaknesses approach.

III. Cons: not criterion-referenced. Major concerns about testimonial type references

IV. Suggestions for report

A. mixed feelings about descriptions of skills vs. yes/no format

B. need for criterion basis

C. beware summative assessment as summation of formative assessment as formative assessments are often more value based. Need to be factual in SA

D. need for evidence, particularly for attitudinal attributes.

E. any difficulty needs to be flagged up early

F. need to use other resources to back up evidence from within training practices

V. Hassles with trainees - yes, clinical care, attitudinal, particularly team work

VI. Conclusions

A. attitudinal criteria - attitudinal problems need to be persistent, interfering with patient care, and evidence from multiple observers.

B. answer format best as yes/no/yes but

C. SA needs to have a temporal element - namely once an element has been passed, it can then be left behind

D. need to identify those elements which are crucial vs. those that are desirable.

SEVEN

I. Current system: mid-term assessment. Principally an assessment of the trainees using MRS, video-tapes, interviews. Some assessment of trainer using video-taped tutorials.

II. Pros: opportunity to address training problems. Opportunity for comments by trainee to external assessor.

III. Cons: rating scales difficult to use.

IV. Suggestions for report:

A. avoid duplication

B. need method for appeal if fail

C. need for second opinion if failure a possibility

D. base on criteria rather that peer reference

E. need specific instructions for exactly what criteria mean

F. suggest yes/no/yes, but , with space for comments and evidence

G. offer the possibility of devolving assessment to other members of the team

V. Hassles with trainees - yes, attitudinal

VI. Conclusions: report needs to be short. Layout easy to read. Area for comments. Separate essential criteria from desirable.

EIGHT

I. Current system: formative assessment only. Formal system (West Midlands Regional Formative Assessment Package).

II. Pros: formative assessment, not summative

III. Cons: no summative element.

IV. Suggestions for report

A. No failure without reference to others

B. Yes/no answer format

C. Gps are making the assessment, but evidence from consultants could be used

V. Hassles with trainees - yes; (7/15); two had concerns about knowledge; all others were attitudinal. Report may have helped esp. in provision of more structured feedback. Unlikely to help decision-making process.

NINE

I. Current system: formative assessment only. Formal system (West Midlands Regional Formative Assessment Package).

II. Pros: formative assessment, not summative

III. Cons: no summative element.

IV. Suggestions for report

A. Needs to be explicit from early on e.g. at interview

B. Yes/no answer format

C. Criteria may needs to include basic attitudes e.g. ageist, racist, sexist

D. Consider use of consultants' evidence

V. Hassles with trainees - yes; (8/13); all were attitudinal. Report may have helped esp. in provision of more structured feedback. Unlikely to help decision-making process.

TEN

I. Current system: formative assessment only. Formal system at 3 and 9 months. Currently virtually all trainees take and pass the MRCGP

II. Pros: formative assessment, not summative. Very high pass rate in MRCGP - probably reflects the selection process; training still currently oversubscribed in N Ireland

III. Cons: nil identified

IV. Suggestions for report

A. Suggest use of yes/no/?don't know format

B. Automatic right of appeal for trainees

C. No "no" without consultation with others

D. Criterion based

E. Need well specified minimum standards

F. Prefer form with standard/evidence/yes/no all together for each element

V. Hassles with trainees - yes; (1/15); global K/S/A. Unsure if structured report would have helped

ELEVEN

I. Current system: formative assessment only. Formal system

II. Pros: formative assessment, not summative. Not prescriptive

III. Cons: no summative element.

IV. Suggestions for report

A. Use of specific guidelines

B. Yes/no answer format

C. Trainees to be aware of contents from time of appointment

D. Wish to emphasise use of consultants to sign up for clinical skills

V. Hassles with trainees - yes; (1/10); one trainer, two episodes - both were global concerns.

TWELVE

I. Current system: formative assessment using MRS, PEP, and OSCE. Summative system - not very formal, and based on MRS.

II. Pros: not discussed in detail

III. Cons: MRS do not cover sufficient areas. Rating scales not very satisfactory.

IV. Suggestions for report

A. Use of evidence

B. Not rating scales format

C. Trainees to be aware of contents from time of appointment

D. Wish to emphasise use of consultants to sign up for clinical skills, preferably before arrival on scheme

E. Each item to be judged individually

V. Hassles with trainees - yes; (2/8); attitudinal; confidence. Felt that structured report could have helped

THIRTEEN

I. Current system: formative assessment using MRS and other systems. Have attempted to draw up a trainer's report - started by developing a core curriculum.

II. Pros: SA - core curriculum based

III. Cons: MRS not very good - the modified areas cover aspects that are too broad. Rating scales not very satisfactory - usual problems with halo and other effects.

IV. Suggestions for report

A. Not MRS

B. Need for minimal standards

C. Avoid grading - go for simple pass/fail

D. Keep the document short

E. Keep the document simply-phrased/jargon-free

V. Hassles with trainees - yes; (5/12); 4 attitudinal; 1 concerns re illness. Structured report would have helped because of the use of specified standards

APPENDIX 5.2:

Draft trainer's report resulting from studies 1-4

SUMMATIVE ASSESSMENT - THE TRAINER'S REPORT

You are being asked to assess the trainee and to indicate your assessment as to ***whether or not the trainee has reached the standard for independent general practice.***

On the next two pages guidance is given about the completion of the report. Please read this guidance carefully; if you have any doubts about the report please consult with your Course Organiser/Associate Adviser.

In this booklet you will find consensus *minimal* standards for each item to help you when you make your decision. It also provides information on the evidence you will need to seek in order to complete this report. Within the report space is provided for you to document the type of evidence used, and when it was undertaken.

To complete this component of summative assessment successfully the trainee needs to have reached the standard for independent general practice for ***all items***. If you decide that for any item the trainee has not reached that standard, please supply details of the evidence on which that decision is based; this will need to include records of the events on which the decision is based, records of discussions that you have undertaken with others involved in the training of this trainee (in accordance with your Regional policy), and records of discussions you have held with the trainee warning the trainee that failure on this item is likely. Space is provided at the end of the report for these records.

GUIDANCE FOR COMPLETING THE TRAINER'S REPORT

Background: The aim of summative assessment is to identify those trainees who are not ready for independent general practice. There are four components to the summative assessment package - a multiple choice questionnaire, an analysis of consultation skills, a submission of written work, and a trainer's report. The trainee needs to pass all four components in order to pass summative assessment. This guidance is provided to assist with the completion of the trainer's report.

What to assess: The contents of the trainer's report are based on the results of a national survey of the views of trainers as to what should be included within it. The report is divided into six sections: "patient care" (itself divided into general clinical skills, patient management skills, and clinical judgement), "communication skills", "organisational skills", "professional values", "personal and professional growth" and "specific clinical skills". The sub-section entitled "specific clinical skills" includes a number of basic diagnostic and therapeutic skills. **Because the trainee needs to pass all items, all items will need to have been tested.**

How to assess: For most trainees the issue of pass/fail will not be a problem; it will be clear that they reach the standard for independent practice as represented by the basic statement given for each item.

However a small proportion of trainees will not be fit for independent practice. In order to help decide whether or not a trainee is fit to practice this document lists minimal standards for each item under the basic statement for each item. These standards were produced by a consensus group of experienced trainers. Each minimal standard is given in the form "what would constitute a failure?"; often more than one standard is given - although it is not required that you test every individual minimal standard, **evidence of a failure for any one of the standards would constitute a failure for this item, and a failure of any item would in turn mean that the trainer's report had been failed.**

To assess the items you will need to gather evidence about the trainee. Evidence can be sought by three main methods:

1. The best evidence is direct observation of the trainee (by sitting in with the trainee or using video-taped recordings); standards assessible in this way are marked (1). Whenever possible, evidence should be collected in this way.
2. For some of the standards tutorial-based discussions may be suitable (for example, problem or random case analysis, case discussion); the standards for which such methods might be suitable are marked (2).
3. Occasionally specific methods might be suitable; the standards for which these methods might be suitable are marked (3), and the specific methods are listed under the title "evidence - specific methods". Some of these methods (particularly OSCE and simulated surgeries) should be undertaken in conjunction with other trainers and Course Organisers/Associate Advisers, whilst others are suitable for assessment within the practice.

Interpreting the standards: For most of the standards the terms "repeatedly" or "persistently" are used. These terms are used for two reasons. Firstly, what is of most concern is unsatisfactory performance that is *likely to continue* once the trainee enters independent practice; this is most likely to happen if it has been seen to happen repeatedly during the training year. Secondly, trainees should not be failed on the basis of a single chance (we are all allowed to make mistakes). **Thus whenever there is any doubt about whether or not the trainee has reached the necessary standard repeated observations should have been made.**

For the items included in the **"specific clinical skills"** subsection, it is recognised that many of these skills may have been tested at a very basic level prior to qualification as a doctor, **it should be emphasised that the requirement in this report is for an assessment with a view to *independent general practice*. The minimal standards therefore not only include standards about the ability to undertake the skills, but also standards about the interpretation of findings.** It will usually be possible to judge whether or not the trainee can undertake the skill successfully by observing the trainee

once (although if the trainee is not able to undertake the skill it will be necessary for the observations to be repeated until the observer is happy that the trainee can undertake the skill successfully); when judging whether or not the trainee can interpret the findings made, it will be necessary for the interpretation of findings to be judged on a number of occasions to ensure that the trainee is interpreting findings in a reliable way; observation on one occasion will not be sufficient. Please remember that ultimately it is your judgement that counts in the completion of the trainer's report; you need to judge whether or not the trainee has reached the stated standard.

If you are considering failing the trainee on an item in the report you should discuss your concerns with your local group of trainers to ensure that your interpretation of the standards fits with the consensus of other trainers. If you are still in doubt, discuss it with the Regional Adviser.

Who to assess: When collecting evidence, people other than yourself may be able to supply suitable evidence (i.e. evidence of the type described above). When other sources are appropriate, these are listed under the title "sources other than trainer". When relying on assessments made by others you need to be sure that they have used one of the methods described above. For a number of standards one of the specific methods listed is "patient/carer complaints"; whilst complaints may be particularly relevant for these standards, a substantiated complaint may form important evidence for any of the items.

Again please remember that ultimately it is your judgement that counts in the completion of the trainer's report; whatever evidence you are using from whatever source you must be happy that the evidence is reliable.

When to assess: Whilst assessment can be going on throughout the training year **most of the trainer's report should not be completed until the final two months of the training year** to ensure that the report does actually reflect performance throughout the training year. The one exception is the "specific clinical skills" component which is located towards the end of the report (items 20a-j, 21a-b); items in this section can be completed during the year as each assessment is undertaken.

In order to ensure that the trainee is given the best chance of passing the trainer's report any concerns about the trainee should be highlighted as early as possible. It is recommended that if you are at all concerned that the trainee is not likely to pass all the items of the report by the end of the training year you should discuss it with both the trainee and your Course Organiser/Associate Adviser (who will then inform your Regional Adviser) by the end of the first three months of the training period.

Records of assessment: Because you might have to be able to confirm that an item has been assessed you should record all the assessments made; there is room on the trainer's report to do this. If you have had any concerns about whether or not the trainee should pass any items you should also keep a record of the discussions you have held with the trainee, other trainers, Course Organiser, Associate Adviser, or Regional Adviser; there is space provided at the end of the report for this purpose.

PATIENT CARE - *GENERAL CLINICAL SKILLS*

1: the doctor can recognise common physical psychological and social problems

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

2: the doctor is able to examine each system and each organ proficiently

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

3: the doctor has the knowledge and skills to deal with life events and crises

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

PATIENT CARE: *GENERAL CLINICAL SKILLS*

1: the doctor can recognise common physical psychological and social problems

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to recognise the presentations of common life-threatening illness	1,2,3
doctor repeatedly fails to recognise the patterns of presentation of common physical, psychological or social problems in patients	1,2,3
doctor repeatedly fails to recognise the physical, psychological and social dimensions of presenting problems	1,3

Sources other than trainer	*Specific methods for items marked 3*
partner, course organiser, PHCT members	OSCE, use of standard cases, patient/carer complaints, notes review, review of trainee log

2: the doctor is able to examine each system and each organ proficiently

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor is unable to undertake successfully a comprehensive examination or an important piece of examination	1,3

Sources other than trainer	*Specific methods for items marked 3*
partner, nurses, consultant. Diplomas may be taken into account.	OSCE, check list for each system/organ

3: the doctor has the knowledge and skills to deal with life events and crises

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to recognise or understand the importance of life events and crises to patients	1,2,3
doctor repeatedly fails to respond to life events or crises presented to him/her	1,2,3
doctor repeatedly fails to utilise the resources available to deal with such events (including material, personal or professional resources)	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	role play, patient/carer feedback/complaints, critical incident technique

4: the doctor demonstrates a broad knowledge of all aspects of the appropriate use of drugs (including actions, interactions, side effects, costs and legal aspects)

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

5: the doctor diagnoses and manages acute emergency situations appropriately

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

PATIENT CARE: *PATIENT MANAGEMENT SKILLS*

6: within his/her assessment the doctor includes the patients' beliefs, ideas, concerns, expectations and the effects of the problem

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

4: the doctor demonstrates a broad knowledge of all aspects of the appropriate use of drugs (including actions, interactions, side effects, costs and legal aspects)

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly prescribes inappropriately (including failure to use relevant drugs, failure to use appropriate doses/preparations/quantities, failure to review long-term treatments, having no recognition of potential side-effects or interactions, having no recognition of drug costs)	1,3*
doctor is repeatedly unable to demonstrate a knowledge of drugs he/she prescribes and is unaware of sources of such information	1,2
doctor is persistently unaware of the risks and regulations associated with controlled drugs (including dependency and legal obligations)	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, course organiser, pharmacist, FHSA pharmaceutical adviser	review of prescriptions (inc. CD register, PACT/SPA data), structured interview, notes review(* only)

5: the doctor diagnoses and manages acute emergency situations appropriately

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to diagnose life-threatening emergencies (including obtaining sufficient information, carrying suitable diagnostic equipment)	1,2,3
doctor repeatedly fails to treat life-threatening emergencies appropriately (including carrying suitable emergency drugs, formulating appropriate management plans to include admission/referral when necessary)	1,2,3
doctor repeatedly fails to cope personally with the stress of emergency situations see also item 11	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members, consultant	critical incident technique, "emergency check lists", outcome analysis of on-call notes, BASICS certificate

PATIENT CARE: *PATIENT MANAGEMENT SKILLS*

6: within his/her assessment the doctor includes the patients' beliefs, ideas, concerns, expectations and the effects of the problem

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to seek the patients ideas, concerns, expectations, beliefs and the effects of the problem	1,3
doctor repeatedly fails to take into account the patients ideas, concerns, expectations, beliefs and the effects of the problem	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	patient/carer complaints, notes review

7: the doctor undertakes examination with appropriate consideration of the patients needs and feelings

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

8: the doctor chooses appropriate management for each problem with the patient (including the care of chronic problems)

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

9: the doctor provides appropriate care and support for patients and their families

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

7: the doctor undertakes examination with appropriate consideration of the patients needs and feelings

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly proceeds with examination against the patient's wishes	1,3
doctor repeatedly fails to take account of patient's dignity (including privacy), sensitivities (including gender, age, culture), or discomfort	1,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	patient/carer complaints

8: the doctor chooses appropriate management for each problem with the patient (including the care of chronic problems)

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to manage problems within consistently accepted good clinical practice (in particular, failing to assess appropriately the presenting problems, failing to consider appropriate range of management options, failing to check on drug reactions)	1,2,3
doctor repeatedly fails to practise "patient-centred" medicine (in particular, communicating/negotiating with patients and families, discussion of long-term implications of diagnosis and treatment with the patient)	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	patient/carer complaints, notes review, simulated surgery

9: the doctor provides appropriate care and support for patients and their families

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to recognise the needs of family or carers	1,2,3
doctor is unaware of or repeatedly fails to utilise support agencies (including PHCT members)	1,2,3
doctor repeatedly fails to perceive the impact of illness of members of the patient's family	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members, hospital consultants	patient/carer feedback/complaints, team meetings, trainee log

PATIENT CARE: *CLINICAL JUDGEMENT*

10: the doctor undertakes appropriate examination (including investigations)

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

11: the doctor responds appropriately to requests for urgent attendance at patients

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

PATIENT CARE: *CLINICAL JUDGEMENT*

10: the doctor undertakes appropriate examination (including investigations)

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly neglects to undertake a comprehensive examination or an important piece of examination (including investigation) when indicated	1,3*
doctor repeatedly undertakes unjustified examination	1,3**

Sources other than trainer	*Specific methods for items marked 3*
partner, nurses, consultants	notes review, OSCE(*only), patient/carer complaints(**only)

11: the doctor responds appropriately to requests for urgent attendance at patients

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to attend medical emergencies within a reasonable time (including failure to ensure that he/she is contactable, failure to communicate effectively with the person requesting help, failure to assess the situation appropriately, failure to act appropriately)	1,2,3
doctor has no understanding of what conditions may present urgently or require urgent management	1,2,3*

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	role play, OSCE, patient/carer complaints, response-time audit, telephone log, notes review(*only). Feedback from deputising/cooperative service may be acceptable.

COMMUNICATION SKILLS

12: the doctor demonstrates effective communication skills when dealing with patients

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

PERSONAL AND PROFESSIONAL GROWTH

13: the doctor is able to identify strengths and weaknesses in his/her performance

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

COMMUNICATION SKILLS

12: the doctor demonstrates effective communication skills when dealing with patients

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to create rapport with the patient (including listening, explaining, and noticing patient cues)	1,3
doctor repeatedly fails to clarify the patient's reason for consulting	1,3
doctor repeatedly fails to convey information to the patient on his/her assessment and management plan that enables the patient to understand what is being said (including the use of language tailored to the particular patient)	1,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	role play, OSCE, simulated surgery, patient/carer complaints

PERSONAL AND PROFESSIONAL GROWTH

13: the doctor is able to identify strengths and weaknesses in his/her performance

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor makes the same mistakes repeatedly to the detriment of patients and is unable to recognise problems within himself/herself that lead to these mistakes	1,2,3
doctor is persistently unable or unwilling to change his/her behaviour to prevent such mistakes when the causes are made known to him/her	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, course organiser, PHCT members	patient/carer complaints, audit, logbooks

ORGANISATIONAL SKILLS

14: the doctor is aware of his/her own limitations, the skills of others, and the ability to refer or delegate appropriately

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

15: the doctor is able to manage his/her own time

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

16: the doctor understands the obligations of a general practitioner according to the NHS contract and regulations

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

ORGANISATIONAL SKILLS

14: the doctor is aware of his/her own limitations, the skills of others, and the ability to refer or delegate appropriately

MINIMAL STANDARDS - what would constitute a failure?	Evidence
the doctor's assessment of his/her own limitations is persistently different from the assessment made by others, with the result that their own limitations are not recognised	1,2,3*
doctor repeatedly fails to recognise or utilise appropriately the skills of others (in particular other PHCT members, hospitals, social services) resulting in a marked over- or under-use of these services	1,2,3**

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members, hospital consultant	patient/carer complaints, team member complaints, audit, confidence/modified Manchester rating scales(* only), notes review(** only)

15: the doctor is able to manage his/her own time

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor is repeatedly late (including starting surgeries, starting tutorials, completing administration) to a level that causes persistent difficulty for others	1,3*
doctor is persistently inflexible (including the management of urgent calls, the management of complex problems that arise in consultations)	1,2,3
doctor is persistently unable to balance the demands on his/her time (including personal vs. professional demands, priorities within working time)	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members (including reception/office staff)	patient/carer complaints, starting time/waiting time audit(* only)

16: the doctor understands the obligations of a general practitioner according to the NHS contract and regulations

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor demonstrates behaviour that would put himself/herself at risk of disciplinary action by the GMC/ Health Authority/Health Board(or their equivalent within the Armed Forces) or at risk of a civil negligence procedure (in particular negligence (including failure to examine or visit) or unethical behaviour (including prescribing/DDA regulations, record-keeping, fraud, dishonesty))	1,2,3
doctor repeatedly fails to recognise behaviours that would put him/her at risk of disciplinary action	1,2,3
see also item 17	

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members, practice manager	patient/carer complaints, service hearing, notes review

PROFESSIONAL VALUES

17: the doctor possesses and applies ethical principles

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

18: the doctor is able to maintain his/her own physical and mental health to a level which enables him/her to discharge the duties of a general medical practitioner

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

19: the doctor is willing to accept appropriate responsibility for patients, partners, colleagues and others

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

PROFESSIONAL VALUES

17: the doctor possesses and applies ethical principles

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly breaches the accepted codes of professional behaviour to a level that puts him/her at risk of disciplinary action by the GMC (or by the Armed Services) (in particular, confidentiality; sexual behaviour; racial, sexual, or religious discrimination; respect for colleagues; ethics of research)	1,2,3
doctor is persistently unaware of the published ethical guidelines ("Professional Conduct and Discipline: Fitness to Practice" (GMC))	2

Sources other than trainer	*Specific methods for items marked 3*
partner, course organiser, PHCT members, consultants	patient/carer complaints, simulated surgeries, role play

18: the doctor is able to maintain his/her own physical and mental health to a level which enables him/her to discharge the duties of a general medical practitioner

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor has a physical or mental illness or a habit (including addiction to drugs or alcohol) which seriously interferes with the provision of effective clinical practice and which he/she is unable or unwilling to control	1,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members, police, court, previous employers	patient/carer complaints, trainee statement of good health, sickness record, police information

19: the doctor is willing to accept appropriate responsibility for patients, partners, colleagues and others

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to ensure that patient care is not prejudiced by his/her absence without good reason (including visits, surgeries), by failure to communicate with others, or by limitations of his/her own performance (including failure to refer, failure to follow-up)	1,2,3
doctor repeatedly abuses patients, staff or colleagues (including verbal, physical or psychological abuse)	1
doctor repeatedly violates his/her contract of employment	1

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	notes review, patient/carer complaints

SPECIFIC CLINICAL SKILLS: *Diagnostic skills*

20. The doctor is able to undertake the following aspects of examination proficiently AND to interpret the findings made:

a): the mental state

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

b): the auroscope

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

c): the ophthalmoscope

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

SPECIFIC CLINICAL SKILLS: *Diagnostic skills*

20. The doctor is able to undertake the following aspects of examination proficiently AND to interpret the findings made:

a): the mental state

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor is repeatedly unable to take a mental health history/examination that allows identification of risks of harm to patient or others (in particular, depression with suicidal ideation/intent, psychoses, confusional states)	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, CPN, trained counsellor, consultant psychiatrist, possession of MRCPsych	OSCE, notes review, simulated surgeries

b): the auroscope

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor persistently fails to maintain equipment in good working order	1,3*
doctor is repeatedly unable to examine the ear (including preparing patients of all ages, visualising the external auditory meatus and tympanic membrane)	1,3**
doctor repeatedly fails to recognise common ear complaints	1,2,3**
see also item 7	

Sources other than trainer	*Specific methods for items marked 3*
partner, trained practice nurses, ENT consultant	review of equipment(* only), OSCE(** only), notes review(** only)

c): the ophthalmoscope

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor persistently fails to maintain equipment in good working order	1,3*
doctor is unwilling to recognise the value of ever using the ophthalmoscope	1,2,3
doctor is repeatedly unable to use the ophthalmoscope to examine the eye (including preparing the patient and the room, visualising the fundus)	1,3**
doctor repeatedly fails to demonstrate an understanding of the limits of his/her competence (including visualising the optic fundus and interpreting the findings made)	1,2

Sources other than trainer	*Specific methods for items marked 3*
partner, consultant	referral letter review, review of equipment(* only), OSCE(** only)

d): the sphygmomanometer

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

e): the stethoscope

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

f): the peak flow meter

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

d): the sphygmomanometer

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor is repeatedly unable to use the sphygmomanometer correctly (including the use of appropriate size cuff; placing of cuff, stethoscope, patient and doctor; distinction between phase IV and V sounds)	1,3
doctor is repeatedly unable to apply the findings to clinical practice	1,2,3*

Sources other than trainer	*Specific methods for items marked 3*
partner, trained practice nurse, hospital physician (specialist registrar grade or above)	correlation of reading with another observer, OSCE, notes review (* only)

e): the stethoscope

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to use the stethoscope correctly (including correct positioning for auscultation of chest, heart, abdomen)	1
doctor repeatedly interprets findings incorrectly	1,3
see also items 7 and 12	

Sources other than trainer	*Specific methods for items marked 3*
partner, consultant	role play, OSCE, simulated surgery, simulated heart sounds

f): the peak flow meter

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor fails to use the correct equipment	1,3
doctor repeatedly fails to teach the patient how to use the meter correctly	1,3
doctor repeatedly fails to use or is unable to use appropriate charts to interpret the results correctly	1,2,3
doctor is repeatedly unable to apply the results to clinical practice	1,2,3*

Sources other than trainer	*Specific methods for items marked 3*
partner, patients, trained practice nurse, consultant	role play, OSCE, notes review(* only)

g): the vaginal examination

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

h): the vaginal speculum

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

i): the cervical smear

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

g): the vaginal examination

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor is unable to undertake bimanual examination	1,3
doctor is unable to describe findings systematically	1,3
doctor repeatedly misinterprets the findings made (including failure to detect signs of major abnormality/illness)	1,3
see also items 7, 12, 20h and 20i	

Sources other than trainer	*Specific methods for items marked 3*
partner, trained nurse, consultant, family planning clinic trainer, patients	family planning certificate

h): the vaginal speculum

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to use a clean speculum of appropriate size and to use gloves	1
doctor repeatedly fails to insert and remove speculum into/from the vagina comfortably	1
doctor repeatedly fails to visualise the cervix	1
doctor repeatedly misinterprets the findings made (including failure to detect signs of major abnormality/illness)	
see also items 7, 12, 20g and 20i	

Sources other than trainer	*Specific methods for items marked 3*
partner, trained nurse, consultant, family planning clinic trainer	

i): the cervical smear

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor has no understanding of the technique required to obtain adequate samples for cervical cytology	1,2,3
doctor fails to use a fresh spatula or brush for each patient	1,3
doctor fails to position the spatula/brush in the cervix correctly	1,3*
doctor fails to put specimen on to slide and fix correctly	1,3*
see also items 7, 12, 20g and 20h	

Sources other than trainer	*Specific methods for items marked 3*
partner, trained nurse, consultant, family planning clinic trainer	family planning certificate, inadequate smear rates(* only)

j): the rectal examination (does not include proctoscopy)

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

SPECIFIC CLINICAL SKILLS: *Emergency care*

21. The doctor is able to undertake the following techniques proficiently:

a): the doctor is able to give an intravenous injection

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

b): the doctor is able to give an intramuscular or subcutaneous injection

Assessment by observation	Assessment by discussion	Assessment by specific methods	Comments

Has the trainee reached the standard for independent general practice? YES☐ NO☐

j): the rectal examination (does not include proctoscopy)

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor is unable to undertake rectal examination	1
doctor repeatedly misinterprets the findings made (including failure to detect signs of major abnormality/illness)	1,2
see also items 7 and 12	

Sources other than trainer	*Specific methods for items marked 3*
partner, consultant	

SPECIFIC CLINICAL SKILLS: *Emergency care*

21. The doctor is able to undertake the following techniques proficiently:

a): the doctor is able to give an intravenous injection

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor fails to check that the drug to be administered is correct (including dose and expiry date)	1
doctor fails to use the appropriate aseptic technique (including needle disposal)	1
doctor repeatedly fails to place the needle within a vein	1
doctor has inadequate knowledge of diagnosis and management of anaphylaxis	1,2
see also items 7 and 12	

Sources other than trainer	*Specific methods for items marked 3*
partner, consultant, nurse qualified to give iv injections	

b): the doctor is able to give an intramuscular or subcutaneous injection

MINIMAL STANDARDS - what would constitute a failure?	Evidence
doctor fails to check that the drug to be administered is correct (including dose and expiry date)	1
doctor fails to use appropriate aseptic technique (including correct sites to be used and needle disposal)	1
doctor has inadequate knowledge of diagnosis and management of anaphylaxis	1,2
see also items 7 and 12	

Sources other than trainer	*Specific methods for items marked 3*
partner, trained nurse	

CONCLUSION

Either:

This trainee, Dr ....., has in my opinion reached the standard for independent general practice in all items in this report.

Signed..... Date.....

Or:

This trainee, Dr ....., has NOT reached the standard for independent general practice in item(s)..... . The evidence on which this decision is based is as follows:

This decision has been discussed with the following people on the following dates:

1. others involved in training:

2. the trainee:

Signed..... Date.....

APPENDIX 6.1:

Final version of the structured trainer's report

SUMMATIVE ASSESSMENT - THE STRUCTURED TRAINER'S REPORT

CONFERENCE OF POSTGRADUATE ADVISERS IN GENERAL PRACTICE - UNIVERSITIES OF THE UNITED KINGDOM

SUMMATIVE ASSESSMENT - THE STRUCTURED TRAINER'S REPORT

Registrar No:

Registrar name:

You are being asked to assess the registrar and to indicate your assessment as to *whether or not the registrar has reached the standard for independent general practice.* To complete this component of summative assessment successfully the registrar needs to have reached the standard for independent general practice for all items. If the registrar has not reached that standard for any item you will need to be able to supply details of the evidence on which that decision is based, including records of the events on which the decision is based, records of discussions you have undertaken with others involved in the training (in accordance with your Regional policy), and records of discussions held with the registrar warning the registrar that failure on this item is likely.

Before completing this report *please read the guidance notes which can be found on the next three pages.* If you have any doubts about the report please consult with your Course Organiser/Associate Adviser.

STATEMENT FROM THE JOINT COMMITTEE ON

POSTGRADUATE TRAINING FOR GENERAL PRACTICE

All trainers should show a copy of the UK Regional Advisers' trainer's report to every trainee (registrar) early in each traineeship and confirm in writing that it is a part of summative assessment of vocational training for general practice. All trainers should ensure that trainees (registrars) have been informed in writing that if trainees (registrars) do not take summative assessment their trainers should be able to justify the means by which their signature was informed when signing the VTR/1 form.

GUIDANCE FOR COMPLETING THE TRAINER'S REPORT

INTRODUCTION:

The aim of summative assessment is to identify those registrars who are ready for independent general practice (i.e. whose performance is above a minimum standard) and those who might benefit from additional training (i.e. whose performance is at or below a minimum standard).

There are four components to the summative assessment package - a multiple choice questionnaire, an analysis of consultation skills, a submission of written work, and a trainer's report. To pass summative assessment the registrar needs to pass all four components - if the registrar is not successful for any component they will be passed on to the Regional referral system for further assessment in that component. To pass the trainer's report the registrar needs to pass all the items contained within it.

If, when all your assessments are complete, you are happy that the registrar's performance reaches the set standards for all items then the trainer's report component can be signed as having been passed. If, however, your evidence makes you concerned that performance for any item is at or below the minimum standards you will then need to refer the registrar into your regional referral system (administered by your Regional Adviser) for further assessment.

USING THIS FORM - UNDERTAKING THE ASSESSMENTS:

This form has two functions. Firstly it contains guidance about what standards you should be using in your assessment of the registrar, about who can help you in undertaking your assessment and about what evidence you should gather. All of these are to be found on the left-hand page of each double-page spread. Secondly it acts as the form on which you should record your assessment. The places to record your assessment are to be found on the right-hand page of each double-page spread.

What to assess: The contents of this trainer's report are based on the results of a national survey of the views of trainers as to what should be included within it. The report is divided into six sections: "patient care" (itself divided into general clinical skills, patient management skills, and clinical judgement), "communication skills", "organisational skills", "professional values", "personal and professional growth" and "specific clinical skills". The sub-section entitled "specific clinical skills" includes a number of basic diagnostic and therapeutic skills. **Because the registrar needs to pass all items, all items will need to have been tested.**

When to assess: Whilst assessment can be going on throughout the training year **most of the trainer's report should not be completed and signed until the beginning of the penultimate month of the total twelve months of general practice training;** this is to ensure that the report does actually reflect performance throughout the training year. The one exception is the "specific clinical skills" component which is located towards the end of the report (items 20a-j, 21a-b); items in this section can be completed during the year as each assessment is undertaken.

In order to ensure that the registrar is given the best chance of passing the trainer's

report the registrar should be shown the standards contained in this report *at the beginning of their training*. Furthermore any concerns about the registrar should be highlighted as early as possible - if you are at all concerned that the registrar is not likely to pass all the items of the report by the end of the training year you should discuss it with both the registrar and your Course Organiser/Associate Adviser (who will then inform your Regional Adviser) as soon as possible, if possible by the end of the first three months of the training period.

How to assess: Each item needs to have been tested. To assess the items you will need to gather evidence about the registrar. Evidence can be sought by three main methods:

1. The best evidence is gained by the assessor directly observing the registrar (by sitting in with the registrar or using video-taped recordings). Whenever possible, evidence should be collected in this way.
2. Sometimes tutorial-based discussions may be suitable (for example, problem or random case analysis, case discussion).
3. Occasionally specific methods might be suitable. Some of these methods (particularly OSCE and simulated surgeries) should be undertaken in conjunction with other trainers and Course Organisers/Associate Advisers, whilst others are suitable for assessment within the practice. For a number of standards one of the specific methods listed is "patient/carer complaints"; whilst complaints may be particularly relevant for these standards, a substantiated complaint may form important evidence for any of the items.

To help you decide whether or not a registrar has reached the standard for independent general practice minimum standards are provided for each item (these are based on the consensus views of a national group of experienced trainers). Additionally for each item advice is given about who else could provide suitable evidence to help you in your decision about that item. These are listed under the title **"sources other than trainer".** When relying on assessments made by others you need to be sure that they have used one of the methods described above. Please remember that ultimately it is your judgement that counts in the completion of the trainer's report; whatever evidence you are using from whatever source you must be happy that the evidence is reliable.

For most registrars it will be clear from the evidence that you have collected that their performance is above the minimum standard. However for some registrars you will need to assess their performance closely to see if they are performing above the minimum standards or not. To assist you in this decision each minimum standard indicates exactly what performance would constitute a failure; furthermore, to help you when the decision is difficult, alongside each individual standard advice is given as to which of the above three methods of assessment would be acceptable in the assessment of that particular standard under the title "evidence" (1 = evidence obtained by direct observation of the registrar by the assessor, 2 = evidence obtained by discussion with the registrar, and 3 = evidence obtained by the specific methods listed under "specific methods for standards marked 3").

Interpreting the standards: For most of the standards the terms "repeatedly" or "persistently" are used. These terms are used for two reasons. Firstly, what is of most concern is unsatisfactory performance that is *likely to continue* once the registrar enters independent practice; this is most likely to happen if it has been seen to happen repeatedly

during the training year. Secondly, registrars should not be failed on the basis of a single chance (we are all allowed to make mistakes). Thus **whenever there is any doubt about whether or not the registrar has reached the necessary standard repeated observations should have been made. Furthermore, if you are considering failing the registrar on an item in the report you must discuss your concerns with others involved in training (for example, your local group of trainers) to ensure that your interpretation of the standards fits with the consensus of other trainers. If you are still in doubt, discuss it with the Regional Adviser.**

For the items included in the "**specific clinical skills**" subsection, it is recognised that many of these skills may have been tested at a basic level prior to qualification as a doctor. **It should be emphasised that the requirement in this report is for an assessment with a view to *independent general practice*. The minimum standards therefore not only include standards about the ability to undertake the skills, but also standards about the interpretation of findings.** It will usually be possible to judge whether or not the registrar can undertake the skill successfully by observing the registrar once (although if the registrar is not able to undertake the skill it will be necessary for the observations to be repeated until the observer is happy that the registrar can undertake the skill successfully), when judging whether or not the registrar can interpret the findings made, it will be necessary for the interpretation of findings to be judged on a number of occasions to ensure that the registrar is interpreting findings in a reliable way; observation on one occasion will not be sufficient. Please remember that ultimately it is <u>your</u> judgement that counts in the completion of the trainer's report; you need to judge whether or not the registrar has reached the stated standard.

<u>USING THIS FORM - RECORDING YOUR ASSESSMENT:</u>

On each right-hand page of the double-page spread there is room for you to make your records about what types of assessment have been used, and whether or not the registrar has reached the standard for independent general practice.

If you keep a separate detailed log of the dates and outcomes of all your assessments then it would be sufficient simply to tick the boxes to indicate the types of assessment used (as in example 1 overleaf). If you do not keep a separate log it is advised that you keep more extensive records of the types, dates, and outcomes of the assessments you have undertaken (as in example 2 overleaf). **If your conclusion is that the registrar does not reach the minimum standards you should keep full records of the types, dates, and outcomes of the assessments you have undertaken (as in example 3 overleaf).**

If you have had any concerns about whether or not the registrar should pass any items you should also keep a record of the discussions you have held with the registrar, other trainers, Course Organiser, Associate Adviser, or Regional Adviser. Space provided at the end of the report for this purpose.

If there are particular comments you want to make about performance for a particular item (for example, comments about particularly good performance or suggestions about improving performance that you might wish to record), please feel free to use the comments section provided for each item.

EXAMPLES OF RECORDS:

1. Minimum record:

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

2. Extensive record:

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

3. Record in case of failure

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

GUIDANCE NOTES

NB under evidence 1=direct observation, 2=evidence from discussion, 3=evidence from specific method listed below

PATIENT CARE: *GENERAL CLINICAL SKILLS*

1: the doctor can recognise common physical, psychological and social problems

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to recognise the presentations of common life-threatening illness	1,2,3
doctor repeatedly fails to recognise the patterns of presentation of common physical, psychological or social problems in patients	1,2,3
doctor repeatedly fails to recognise the physical, psychological and social dimensions of presenting problems	1,3

Sources other than trainer	*Specific methods for items marked 3*
partner, course organiser, PHCT members	OSCE, use of standard cases, patient/carer complaints, notes review, review of registrar log

2: the doctor is able to examine each system and each organ proficiently

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor is unable to undertake successfully a comprehensive examination or an important piece of examination	1,3

Sources other than trainer	*Specific methods for items marked 3*
partner, nurses, consultant. Diplomas may be taken into account.	OSCE, check list for each system/organ

3: the doctor has the knowledge and skills to deal with life events and crises

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to recognise or understand the importance of life events and crises to patients	1,2,3
doctor repeatedly fails to respond to life events or crises presented to him/her	1,2,3
doctor repeatedly fails to utilise the resources available to deal with such events (including material, personal or professional resources)	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	role play, patient/carer feedback/complaints, critical incident technique

RECORDS PAGE

PATIENT CARE - *GENERAL CLINICAL SKILLS*

1: the doctor can recognise common physical, psychological and social problems

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

2: the doctor is able to examine each system and each organ proficiently

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

3: the doctor has the knowledge and skills to deal with life events and crises

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

GUIDANCE NOTES

NB under evidence 1=direct observation, 2=evidence from discussion, 3=evidence from specific method listed below

4: the doctor demonstrates a broad knowledge of all aspects of the appropriate use of drugs (including actions, interactions, side effects, costs and legal aspects)

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly prescribes inappropriately (including failure to use relevant drugs, failure to use appropriate doses/preparations/quantities, failure to review long-term treatments, having no recognition of potential side-effects or interactions, having no recognition of drug costs)	1,3*
doctor is repeatedly unable to demonstrate a knowledge of drugs he/she prescribes and is unaware of sources of such information	1,2
doctor is persistently unaware of the risks and regulations associated with controlled drugs (including dependency and legal obligations)	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, course organiser, pharmacist, FHSA pharmaceutical adviser	review of prescriptions (inc. CD register, PACT/SPA data), structured interview, notes review(* only)

5: the doctor diagnoses and manages acute emergency situations appropriately

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to diagnose life-threatening emergencies (including obtaining sufficient information, carrying suitable diagnostic equipment)	1,2,3
doctor repeatedly fails to treat life-threatening emergencies appropriately (including carrying suitable emergency drugs, formulating appropriate management plans to include admission/referral when necessary)	1,2,3
doctor repeatedly fails to cope personally with the stress of emergency situations **see also item 11**	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members, consultant	critical incident technique, "emergency check lists", outcome analysis of on-call notes, BASICS certificate

PATIENT CARE: *PATIENT MANAGEMENT SKILLS*

6: within his/her assessment the doctor includes the patients' beliefs, ideas, concerns, expectations and the effects of the problem

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to seek the patients ideas, concerns, expectations, beliefs and the effects of the problem	1,3
doctor repeatedly fails to take into account the patients ideas, concerns, expectations, beliefs and the effects of the problem	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	patient/carer complaints, notes review

RECORDS PAGE

4: the doctor demonstrates a broad knowledge of all aspects of the appropriate use of drugs (including actions, interactions, side effects, costs and legal aspects)

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

5: the doctor diagnoses and manages acute emergency situations appropriately

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

PATIENT CARE: *PATIENT MANAGEMENT SKILLS*

6: within his/her assessment the doctor includes the patients' beliefs, ideas, concerns, expectations and the effects of the problem

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

GUIDANCE NOTES

NB under evidence 1=direct observation, 2=evidence from discussion, 3=evidence from specific method listed below

7: the doctor undertakes examination with appropriate consideration of the patients' needs and feelings

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly proceeds with examination against the patient's wishes	1,3
doctor repeatedly fails to take account of patient's dignity (including privacy), sensitivities (including gender, age, culture), or discomfort	1,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	patient/carer complaints

8: the doctor chooses appropriate management for each problem with the patient (including the care of chronic problems)

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to manage problems within consistently accepted good clinical practice (in particular, failing to assess appropriately the presenting problems, failing to consider appropriate range of management options, failing to check on drug reactions)	1,2,3
doctor repeatedly fails to practise "patient-centred" medicine (in particular, communicating/negotiating with patients and families, discussion of long-term implications of diagnosis and treatment with the patient)	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	patient/carer complaints, notes review, simulated surgery

9: the doctor provides appropriate care and support for patients and their families

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to recognise the needs of family or carers	1,2,3
doctor is unaware of or repeatedly fails to utilise support agencies (including PHCT members)	1,2,3
doctor repeatedly fails to perceive the impact of illness of members of the patient's family	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members, hospital consultants	patient/carer feedback/complaints, team meetings, registrar log

RECORDS PAGE

7: the doctor undertakes examination with appropriate consideration of the patients' needs and feelings

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

8: the doctor chooses appropriate management for each problem with the patient (including the care of chronic problems)

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

9: the doctor provides appropriate care and support for patients and their families

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

GUIDANCE NOTES

NB under evidence 1=direct observation, 2=evidence from discussion, 3=evidence from specific method listed below

PATIENT CARE: *CLINICAL JUDGEMENT*

10: the doctor undertakes appropriate examination (including investigations)

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly neglects to undertake a comprehensive examination or an important piece of examination (including investigation) when indicated	1,3*
doctor repeatedly undertakes unjustified examination	1,3**

Sources other than trainer	*Specific methods for items marked 3*
partner, nurses, consultants	notes review, OSCE(*only), patient/carer complaints(**only)

11: the doctor responds appropriately to requests for urgent attendance at patients

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to attend medical emergencies within a reasonable time (including failure to ensure that he/she is contactable, failure to communicate effectively with the person requesting help, failure to assess the situation appropriately, failure to act appropriately)	1,2,3
doctor has no understanding of what conditions may present urgently or require urgent management	1,2,3*

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	role play, OSCE, patient/carer complaints, response-time audit, telephone log, notes review(*only). Feedback from deputising/cooperative service may be acceptable.

RECORDS PAGE

PATIENT CARE: *CLINICAL JUDGEMENT*

10: the doctor undertakes appropriate examination (including investigations)

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

11: the doctor responds appropriately to requests for urgent attendance at patients

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

GUIDANCE NOTES

NB under evidence 1=direct observation, 2=evidence from discussion, 3=evidence from specific method listed below

COMMUNICATION SKILLS

12: the doctor demonstrates effective communication skills when dealing with patients

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to create rapport with the patient (including listening, explaining, and noticing patient cues)	1,3
doctor repeatedly fails to clarify the patient's reason for consulting	1,3
doctor repeatedly fails to convey information to the patient on his/her assessment and management plan that enables the patient to understand what is being said (including the use of language tailored to the particular patient)	1,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	role play, OSCE, simulated surgery, patient/carer complaints

PERSONAL AND PROFESSIONAL GROWTH

13: the doctor is able to identify strengths and weaknesses in his/her performance

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor makes the same mistakes repeatedly to the detriment of patients and is unable to recognise problems within himself/herself that lead to these mistakes	1,2,3
doctor is persistently unable or unwilling to change his/her behaviour to prevent such mistakes when the causes are made known to him/her	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, course organiser, PHCT members	patient/carer complaints, audit, logbooks

RECORDS PAGE

COMMUNICATION SKILLS

12: the doctor demonstrates effective communication skills when dealing with patients

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

PERSONAL AND PROFESSIONAL GROWTH

13: the doctor is able to identify strengths and weaknesses in his/her performance

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

GUIDANCE NOTES

NB under evidence 1=direct observation, 2=evidence from discussion, 3=evidence from specific method listed below

ORGANISATIONAL SKILLS

14: the doctor is aware of his/her own limitations, the skills of others, and the ability to refer or delegate appropriately

MINIMUM STANDARDS - what would constitute a failure?	Evidence
the doctor's assessment of his/her own limitations is persistently different from the assessment made by others, with the result that their own limitations are not recognised	1,2,3*
doctor repeatedly fails to recognise or utilise appropriately the skills of others (in particular other PHCT members, hospitals, social services) resulting in a marked over- or under-use of these services	1,2,3**

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members, hospital consultant	patient/carer complaints, team member complaints, audit, confidence/modified Manchester rating scales(* only), notes review(** only)

15: the doctor is able to manage his/her own time

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor is repeatedly late (including starting surgeries, starting tutorials, completing administration) to a level that causes persistent difficulty for others	1,3*
doctor is persistently inflexible (including the management of urgent calls, the management of complex problems that arise in consultations)	1,2,3
doctor is persistently unable to balance the demands on his/her time (including personal vs. professional demands, priorities within working time)	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members (including reception/office staff)	patient/carer complaints, starting time/waiting time audit(* only)

16: the doctor understands the obligations of a general practitioner according to the NHS contract and regulations

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor demonstrates behaviour that would put himself/herself at risk of disciplinary action by the GMC/ Health Authority/Health Board (or their equivalent within the Armed Forces) or at risk of a civil negligence procedure (in particular negligence (including failure to examine or visit) or unethical behaviour (including prescribing/DDA regulations, record-keeping, fraud, dishonesty))	1,2,3
doctor repeatedly fails to recognise behaviours that would put him/her at risk of disciplinary action **see also item 17**	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members, practice manager	patient/carer complaints, service hearing, notes review

RECORDS PAGE

ORGANISATIONAL SKILLS

14: the doctor is aware of his/her own limitations, the skills of others, and the ability to refer or delegate appropriately

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

15: the doctor is able to manage his/her own time

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

16: the doctor understands the obligations of a general practitioner according to the NHS contract and regulations

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

GUIDANCE NOTES

NB under evidence 1=direct observation, 2=evidence from discussion, 3=evidence from specific method listed below

PROFESSIONAL VALUES

17: the doctor possesses and applies ethical principles

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly breaches the accepted codes of professional behaviour to a level that puts him/her at risk of disciplinary action by the GMC (or by the Armed Forces) (in particular, confidentiality; sexual behaviour; racial, sexual, or religious discrimination; respect for colleagues; ethics of research)	1,2,3
doctor is persistently unaware of the published ethical guidelines ("Good Medical Practice" (GMC, 1995))	2

Sources other than trainer	*Specific methods for items marked 3*
partner, course organiser, PHCT members, consultants	patient/carer complaints, simulated surgeries, role play

18: the doctor is able to maintain his/her own physical and mental health to a level which enables him/her to discharge the duties of a general medical practitioner

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor has a physical or mental illness or a habit (including addiction to drugs or alcohol) which seriously interferes with the provision of effective clinical practice and which he/she is unable or unwilling to control	1,3

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members, police, court, previous employers	patient/carer complaints, registrar statement of good health, sickness record, police information

19: the doctor is willing to accept appropriate responsibility for patients, partners, colleagues and others

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to ensure that patient care is not prejudiced by his/her absence without good reason (including visits, surgeries), by failure to communicate with others, or by limitations of his/her own performance (including failure to refer, failure to follow-up)	1,2,3
doctor repeatedly abuses patients, staff or colleagues (including verbal, physical or psychological abuse)	1
doctor repeatedly violates his/her contract of employment	1

Sources other than trainer	*Specific methods for items marked 3*
partner, PHCT members	notes review, patient/carer complaints

RECORDS PAGE

PROFESSIONAL VALUES

17: the doctor possesses and applies ethical principles

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

18: the doctor is able to maintain his/her own physical and mental health to a level which enables him/her to discharge the duties of a general medical practitioner

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

19: the doctor is willing to accept appropriate responsibility for patients, partners, colleagues and others

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

GUIDANCE NOTES

NB under evidence 1=direct observation, 2=evidence from discussion, 3=evidence from specific method listed below

SPECIFIC CLINICAL SKILLS: *Diagnostic skills*

20. The doctor is able to undertake the following aspects of examination proficiently AND to interpret the findings made:

a): the mental state

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor is repeatedly unable to take a mental health history/examination that allows identification of risks of harm to patient or others (in particular, depression with suicidal ideation/intent, psychoses, confusional states)	1,2,3

Sources other than trainer	*Specific methods for items marked 3*
partner, CPN, trained counsellor, consultant psychiatrist, possession of MRCPsych	OSCE, notes review, simulated surgeries

b): using the auriscope

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor persistently fails to maintain equipment in good working order	1,3*
doctor is repeatedly unable to examine the ear (including preparing patients of all ages, visualising the external auditory meatus and tympanic membrane)	1,3**
doctor repeatedly fails to recognise common ear complaints	1,2,3**
see also item 7	

Sources other than trainer	*Specific methods for items marked 3*
partner, trained practice nurses, ENT consultant	review of equipment(* only), OSCE(** only), notes review(** only)

c): using the ophthalmoscope

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor persistently fails to maintain equipment in good working order	1,3*
doctor is unwilling to recognise the value of ever using the ophthalmoscope	1,2,3
doctor is repeatedly unable to use the ophthalmoscope to examine the eye (including preparing the patient and the room, visualising the fundus)	1,3**
doctor repeatedly fails to demonstrate an understanding of the limits of his/her competence (including visualising the optic fundus and interpreting the findings made)	1,2

Sources other than trainer	*Specific methods for items marked 3*
partner, consultant	referral letter review, review of equipment(* only), OSCE(** only)

RECORDS PAGE

SPECIFIC CLINICAL SKILLS: *Diagnostic skills*

20. The doctor is able to undertake the following aspects of examination proficiently AND to interpret the findings made:

a): the mental state

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

b): using the auriscope

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

c): using the ophthalmoscope

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

GUIDANCE NOTES

NB under evidence 1=direct observation, 2=evidence from discussion, 3=evidence from specific method listed below

d): using the sphygmomanometer

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor is repeatedly unable to use the sphygmomanometer correctly (including the use of appropriate size cuff; placing of cuff, stethoscope, patient and doctor; distinction between phase IV and V sounds)	1,3
doctor is repeatedly unable to apply the findings to clinical practice	1,2,3*

Sources other than trainer	*Specific methods for items marked 3*
partner, trained practice nurse, hospital physician (specialist registrar grade or above)	correlation of reading with another observer, OSCE, notes review (* only)

e): using the stethoscope

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to use the stethoscope correctly (including correct positioning for auscultation of chest, heart, abdomen)	1
doctor repeatedly interprets findings incorrectly	1,3
see also items 7 and 12	

Sources other than trainer	*Specific methods for items marked 3*
partner, consultant	role play, OSCE, simulated surgery, simulated heart sounds

f): using the peak flow meter

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor fails to use the correct equipment	1,3
doctor repeatedly fails to teach the patient how to use the meter correctly	1,3
doctor repeatedly fails to use or is unable to use appropriate charts to interpret the results correctly	1,2,3
doctor is repeatedly unable to apply the results to clinical practice	1,2,3*

Sources other than trainer	*Specific methods for items marked 3*
partner, patients, trained practice nurse, consultant	role play, OSCE, notes review(* only)

RECORDS PAGE

d): using the sphygmomanometer

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

e): using the stethoscope

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

f): using the peak flow meter

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

GUIDANCE NOTES

NB under evidence 1=direct observation, 2=evidence from discussion, 3=evidence from specific method listed below

g): the vaginal examination

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor is unable to undertake bimanual examination	1,3
doctor is unable to describe findings systematically	1,3
doctor repeatedly misinterprets the findings made (including failure to detect signs of major abnormality/illness)	1,3
see also items 7, 12, 20h and 20i	

Sources other than trainer	*Specific methods for items marked 3*
partner, trained nurse, consultant, family planning clinic trainer, patients	family planning certificate, DFFP

h): using the vaginal speculum

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor repeatedly fails to use a clean speculum of appropriate size and to use gloves	1
doctor repeatedly fails to insert and remove speculum into/from the vagina comfortably	1
doctor repeatedly fails to visualise the cervix	1
doctor repeatedly misinterprets the findings made (including failure to detect signs of major abnormality/illness)	
see also items 7, 12, 20g and 20i	

Sources other than trainer	*Specific methods for items marked 3*
partner, trained nurse, consultant, family planning clinic trainer	

i): the cervical smear

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor has no understanding of the technique required to obtain adequate samples for cervical cytology	1,2,3
doctor fails to use a fresh spatula or brush for each patient	1,3
doctor fails to position the spatula/brush in the cervix correctly	1,3*
doctor fails to put specimen on to slide and fix correctly	1,3*
see also items 7, 12, 20g and 20h	

Sources other than trainer	*Specific methods for items marked 3*
partner, trained nurse, consultant, family planning clinic trainer	family planning certificate, DFFP, inadequate smear rates(* only)

RECORDS PAGE

g): the vaginal examination

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

h): using the vaginal speculum

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

i): the cervical smear

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

GUIDANCE NOTES

NB under evidence 1=direct observation, 2=evidence from discussion, 3=evidence from specific method listed below

j): the rectal examination (does not include proctoscopy)

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor is unable to undertake rectal examination	1
doctor repeatedly misinterprets the findings made (including failure to detect signs of major abnormality/illness)	1,2
see also items 7 and 12	

Sources other than trainer	*Specific methods for items marked 3*
partner, consultant	

SPECIFIC CLINICAL SKILLS: *Emergency care*

21. The doctor is able to undertake the following techniques proficiently:

a): the doctor is able to give an intravenous injection

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor fails to check that the drug to be administered is correct (including dose and expiry date)	1
doctor fails to use the appropriate aseptic technique (including needle disposal)	1
doctor repeatedly fails to place the needle within a vein	1
doctor has inadequate knowledge of diagnosis and management of anaphylaxis	1,2
see also items 7 and 12	

Sources other than trainer	*Specific methods for items marked 3*
partner, consultant, nurse qualified to give iv injections	

b): the doctor is able to give an intramuscular or subcutaneous injection

MINIMUM STANDARDS - what would constitute a failure?	Evidence
doctor fails to check that the drug to be administered is correct (including dose and expiry date)	1
doctor fails to use appropriate technique (including correct sites to be used, aseptic technique and needle disposal)	1
doctor has inadequate knowledge of diagnosis and management of anaphylaxis	1,2
see also items 7 and 12	

Sources other than trainer	*Specific methods for items marked 3*
partner, trained nurse	

RECORDS PAGE

j): the rectal examination (does not include proctoscopy)

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

SPECIFIC CLINICAL SKILLS: *Emergency care*

21. The doctor is able to undertake the following techniques proficiently:

a): the doctor is able to give an intravenous injection

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

b): the doctor is able to give an intramuscular or subcutaneous injection

Assessment by direct observation of registrar by assessor	Assessment by discussion between registrar and assessor	Assessment by specific methods	Comments

Has the registrar reached the standard for independent general practice? YES☐ NO☐

CONCLUSION

Registrar no: **Registrar name:**

***Either*:**

This registrar, Dr ....., has in my opinion reached the standard for independent general practice in all items in this report.

Signed..... Date.....

NAME OF TRAINER (please print or stamp):

***Or*:**

This registrar , Dr ....., has NOT reached the standard for independent general practice in item(s)..... . The evidence on which this decision is based is as follows:

This decision has been discussed with the following people on the following dates:

1. **others involved in training:**

2. **the registrar:**

Signed..... Date.....

NAME OF TRAINER (please print or stamp):

Checked by: **Action:**

Director of Postgraduate General Practice Education